Sponsoring Committee: Professor Tara McAllister, Chairperson
Professor Adam Buchwald
Professor Douglas H. Whalen

ESTABLISHING THE ROLE OF SENSORIMOTOR SKILLS IN SPEECH DEVELOPMENT AND DISORDERS

Heather Michelle Kabakoff

Program in Communicative Sciences and Disorders
Department of Communicative Sciences and Disorders

Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in the
School of Culture, Education, and Human Development
New York University
2021

ACKNOWLEDGMENTS

My highest degree of gratitude goes to my advisor, Tara McAllister, for her immeasurable guidance over the last eight years. Since the day I began working as a clinician for her biofeedback treatment study, I have transformed from a clinician with an interest in research into an research scientist with a focus on clinical questions. Tara enthusiastically supported my research interests while generously inviting me into her multi-site collaborations, encouraging my independent study of data science, guiding my pursuit of national grants and fellowships, and providing unlimited feedback and advice that always improved the quality of my drafts. I am forever appreciative to her for her sustained commitment to my growth.

Carrying out my dissertation is forever linked to my funding sources, which include a Ruth L. Kirschstein National Research Service Award Individual Predoctoral Fellowship (F31) from the National Institutes of Health National Institute of Deafness and other Communication Disorders, a Stetson Award from the Acoustical Society of America, and a New Century Scholars Doctoral Scholarship from the American Speech-Language-Hearing Foundation. Infinite gratitude goes to my dissertation committee members including my grant co-sponsor Douglas H. Whalen, who taught me articulatory phonetics while providing me with my dissertation data, and grant consultant Adam Buchwald, who guided me through many foundational courses on speech perception and

production. Endless gratitude goes to my grant collaborators and consultants (Daphna Harel, Jonathan Preston, Mark Tiede, Douglas Shiller, Steven Lulich), my mentorship team (Susannah Levi, Elaine Hitchcock, Erika Levy, Yana Yunusova), and my dissertation readers (Maria Grigos, Lisa Davidson).

At New York University, I was fortunate to be a part of Tara McAllister's *Biofeedback Intervention Technology for Speech* Lab and Susannah Levi's *Acoustic Phonetics and Perception* Lab. At the City University of New York, I was grateful to be welcomed into Douglas H. Whalen's *Speech Production, Acoustics, and Perception Laboratory*. I thank the dozens of students and colleagues in these labs for their support and conference companionship.

My family and friends have not only helped me survive the past years of PhD work, they have actually helped me thrive. My mom has continued to drop everything for a close read of my proposal drafts, taking a deep level of interest in my work through her proofreading talent. Thank you to a handful of friends for their persistent interest in my work, and to everyone else for providing me with a life outside of academics. My husband Stu has played a unique role in that he can glance at any type of code that I am struggling with (R, Matlab, or whatever) and skillfully offer suggestions to help me debug and "decomplect" (as he calls it). He also has an eye for design, which has helped encourage my aesthetic to decrease in awkwardness and increase in elegance. He and I were thrilled to welcome baby Elvis Bernard Kabakoff just two weeks after I defended this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION TO DISSERTATION

Speech sound disorder (SSD) is a condition that affects the speech output of approximately one-sixth of preschool-aged children (T. F. Campbell et al., 2003). Children with misarticulated speech sounds experience negative social and emotional challenges that may impact their ability to participate fully in academic and occupational settings later in adulthood (Felsenfeld, Broen, & McGue, 1994). Most children recover either spontaneously or through targeted therapy, but an estimated 25% of children with SSD persist with errors past six years of age (Shriberg, Tomblin, & McSweeny, 1999). This amounts to an approximate 4% of the population who presents with a moderate to severe case of SSD at age six (Shriberg et al., 1999). For the 1-2% of individuals who develop residual speech errors (RSE) that continue into adolescence and adulthood (Flipsen, 2015), the socioemotional challenges they endure may become life-long personal and professional obstacles (Hitchcock, Harel, & McAllister Byun, 2015). Knowing which factors predict who will persist with errors beyond childhood is a crucial first step toward making evidence-based assessment and treatment decisions for this clinical population.

According to the Speech Disorders Classification System (SDCS, Shriberg et al., 2010), SSD has three etiological branches: speech delay, speech errors, and motor speech disorder (*see Figure 1.1*). The speech delay branch includes error patterns with cognitive-linguistic (e.g., genetic), auditory-perceptual (e.g., otitis media), or psycho-social origins. The speech errors branch includes children who have habituated rhotic and sibilant errors with no known etiology. The motor

speech disorder branch includes children whose speech errors are associated with deficits in motor execution. According to the newly revised version of the SDCS, the categories within the Motor Speech Disorders branch include apraxia of speech, dysarthria, a combination of the two, and speech motor delay (Shriberg et al., 2019).

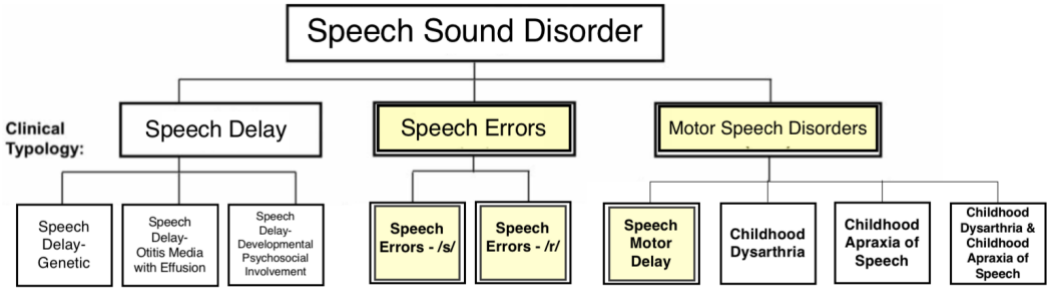

Figure 1.1. Speech Disorders Classification System, adapted from Shriberg, Kwiatkowski, and Mabie (2019) with permission.

Of critical relevance to this dissertation, the diagnostic markers that differentiate isolated speech errors from those associated with speech motor delay have not been directly studied. Within the motor speech disorder branch, neuroimaging can reveal the presence of neural abnormalities associated with dysarthria (Liégeois & Morgan, 2012), while core speech characteristics guide diagnosis of apraxia of speech (Murray, McCabe, Heard, & Ballard, 2015). However, there is no evidence that neuroimaging can be used to distinguish speech motor delay from isolated speech errors. As SSD with typical speech motor control is more likely to resolve than SSD with atypical speech motor control (Vick et al., 2014), it is crucial to be able to differentiate these two clinical subpopulations in young children by identifying those children with motor delays.

In the current study, we develop and test a direct measure of motor execution skill intended to identify motor factors associated with difficulty developing adultlike speech. If the contribution of motor execution skill can be quantified as distinct from other factors influencing motor skill, then children with primarily motor-based impairments could be selected and matched with a corresponding motor-based treatment approach, potentially reducing the duration of treatment required before discharge.

Sensorimotor underpinnings of speech

Among various models of speech production, the Directions Into Velocities of Articulators (DIVA; Guenther, 2016) model provides one theoretical framework for how motor plans are implemented, updated, and executed to produce speech (for similar models, see, e.g., Hickok, 2012; Houde & Nagarajan, 2011; Parrell, Ramanarayanan, Nagarajan, & Houde, 2019). As seen in *Figure 1.2*, the DIVA model posits that skilled speech production involves a feedforward channel (left) and feedback loops (right) that direct articulator placement. During speech production, stored mental representations are translated into auditory and somatosensory targets that determine the feedforward motor plan and how it will be updated through the corresponding feedback channels. The current focus is on the figure's highlighted areas (somatosensory/articulatory maps), which are relatively understudied. Based on this model and on other models of speech production, individual differences in production accuracy may be connected with individual differences in motor execution skill, as well as in auditory and somatosensory acuity.

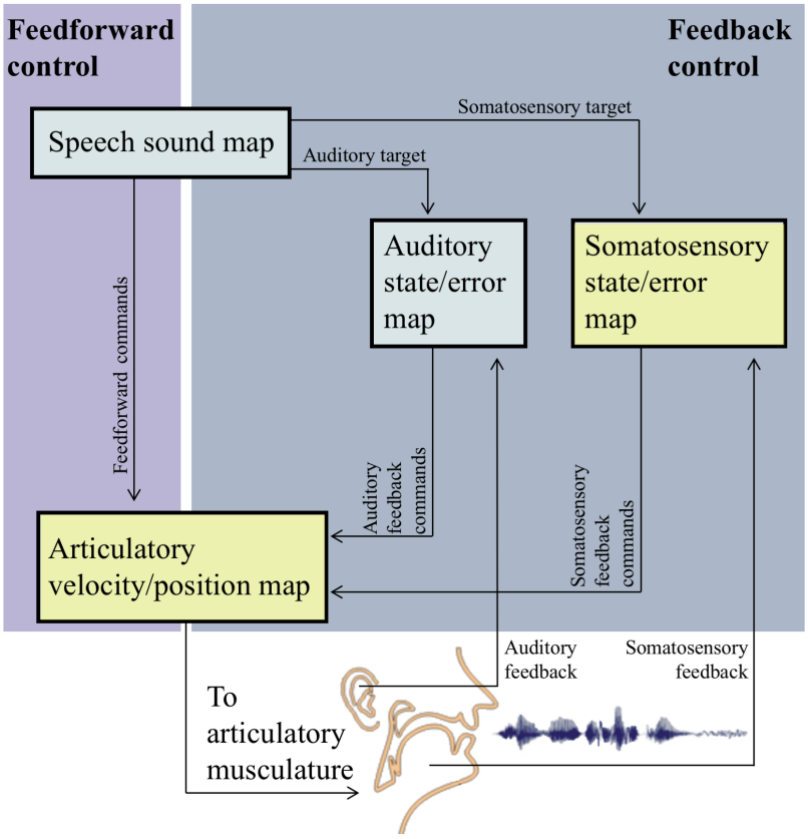

Figure 1.2. Simplified representation of DIVA model. Adapted from Guenther (2016) with permission.

Auditory acuity is known to be correlated with degree of production accuracy in typically developing (TD) children (McAllister Byun & Tiede, 2017) and in adults (Perkell et al., 2004). Relatedly, auditory acuity is also known to be relatively low in children with SSD (Cabbage, Hogan, & Carrell, 2016; Hearnshaw, Baker, & Munro, 2018, 2019; Rvachew & Jamieson, 1989; Rvachew, Ohberg, Grawburg, & Heyding, 2003; Shiller, Rvachew, & Brosseau-Lapré, 2010). Such findings have motivated a line of research focusing on providing input-oriented treatment for those with auditory-perceptual delays (Jamieson & Rvachew, 1992; Rvachew, 1994; Rvachew & Brosseau-Lapré, 2010; Rvachew, Nowak, & Cloutier, 2004).

The roles of somatosensory acuity and motor execution skill in achieving adultlike speech are the focus of the present study and will be addressed in separate sections below. Broadly, research has indicated that an individual's degree of somatosensory acuity represents the degree of refinement of somatosensory targets in combination with the ability to access and respond to somatosensory feedback (Ghosh et al., 2010). Degree of somatosensory acuity is associated with the precision of phonetic output in both typical (Ghosh et al., 2010; Tremblay, Shiller, & Ostry, 2003) and disordered populations (Fucci, 1972; Fucci & Robertson, 1971; McNutt, 1977). In the motor domain, research has suggested that an impaired feedforward mechanism is a primary contributor to the speech error patterns associated with apraxia of speech (Maas, Mailend, & Guenther, 2015), a finding that may extend more broadly to SSD (Terband, Maassen, Guenther, & Brumberg, 2014). For individuals with motor impairment, it follows that optimal treatment should involve repetitive practice to build from producing target sounds in isolation to the conversational level (Van Riper, 1978) while incorporating the principles of motor learning (Maas et al., 2008).

Taken together, if deficiencies in the robustness of the feedforward plan (motor execution skill) and/or in the ability to access and respond to auditory feedback (auditory acuity) or somatosensory feedback (somatosensory acuity) are associated with lack of response to treatment, then treatment should in principle aim to enhance the deficient subsystem. Thus, our theoretically-driven approach aims to measure the degree of refinement that child speakers exert over motor execution, in addition to somatosensory acuity. Our analysis controls for the

better studied covariate of auditory acuity to better understand the joint role of these skills in individuals with SSD.

Measuring motor execution skill

In early stages of speech development (Fletcher, 1989) and in disordered speech (Gibbon, 1999), a reduced ability to independently control anterior versus posterior lingual regions may play a role in children's nonadultlike speech patterns. The ability to isolate movement of different regions of the tongue is referred to as "lingual differentiation," such that gestures lacking a typical degree of independent movement may be described as "undifferentiated gestures." Gibbon (1999) suggested that treatment prognosis may be related with the degree to which a child presents with undifferentiated gestures for targets that require multiple lingual constrictions. In addition, Gibbon (1999) suggested that the presence or absence of widespread undifferentiated gestures could guide the clinician in selecting an intervention approach for a given child. Children who exhibit frequent undifferentiated lingual gestures, suggesting motor involvement, may be best served by a traditional articulatory treatment approach that incorporates the principles of motor learning (Maas et al., 2008). In contrast, children who present with speech errors but show a normal degree of lingual differentiation might be better suited for a phonetic-perceptual or a phonological approach to treatment. This dissertation aims to facilitate the measurement of lingual differentiation with the potential long-term application of determining whether treatment selection based on such measurement can enhance outcomes for children with SSD.

Moving past indirect measures of lingual differentiation

Previous studies of lingual differentiation used techniques that reveal what areas of contact the tongue makes with the palate through sensors placed on the palate. Such methods include palatography (e.g., Fletcher, 1989) and electropalatography (EPG, e.g., Gibbon, 1999). Using these methods, researchers have inferred the presence of lingual differentiation and found developmental trends in linguopalatal contact patterns. That is, TD adolescents use differentiated lingual contours more often than TD school-aged children (Fletcher, 1989). Among school-aged children, TD speakers use differentiated lingual contours more often than those with SSD (Gibbon, 1999). Although research measuring palatal contact has been pivotal in determining the time course of motoric development in populations with and without SSD, the tools for the more commonly used EPG require fitting expensive customized palatal prostheses, which is not a feasible approach in the majority of clinical settings. Additionally, palatal contact patterns determined from EPG are typically interpreted as reflecting either a differentiated or an undifferentiated tongue shape, instead of providing a direct continuous measure of degree of lingual differentiation. Therefore, an approach that directly measures degree of lingual differentiation along a continuous scale could help researchers detect fine-grained differences in motor execution skill within and across individuals.

Direct measures of lingual differentiation

In contrast to the inferred tongue shape patterns derived from EPG, ultrasound imaging reveals a continuous midsagittal lingual contour, which can

yield relevant information about tongue shape even when there is no linguopalatal contact. For several decades, ultrasound has been used for treatment of speech errors in various clinical populations, including adolescents and adults with hearing impairment, residual speech errors, and cleft palate (Bacsfalvi, 2010; Bernhardt, Bacsfalvi, Gick, Radanov, & Williams, 2005; Bernhardt, Gick, Bacsfalvi, & Adler-Bock, 2005; Bernhardt, Gick, Bacsfalvi, & Ashdown, 2003; Shawker & Sonies, 1985). Ultrasound is a relatively accessible, affordable, ready-to-use, and minimally-invasive tool that can be used with small children. Klein, McAllister Byun, Davidson, and Grigos (2013) used ultrasound to demonstrate differences in tongue contours among school-aged children with and without SSD producing rhotic targets. Based on blinded raters' visual impression of whether each contour was differentiated or not, the authors found a greater proportion of differentiated gestures in TD children than in children with SSD, and in perceptually correct productions than in perceptually incorrect productions. For the children with SSD in treatment for rhotic errors, they found a greater number of differentiated gestures after treatment than before treatment. These findings demonstrate that it is possible to derive insight into a child's stage of motor development using ultrasound. Furthermore, the use of ultrasound imaging makes it possible to directly quantify the degree of lingual differentiation of a given contour along a continuous scale.

Degree of tongue complexity

We will henceforth use the term "tongue complexity" to refer to a continuous measure of the degree of differentiation of a given ultrasound-

extracted lingual contour. Previous attempts to automate quantification of the degree of tongue complexity have yielded promising results. The modified curvature index (MCI, Dawson, Tiede, & Whalen, 2016) is an averaging technique that integrates the curvature of a tagged lingual contour with respect to the length of the arc, and minimizes the difference between two adjacent points by taking the integral of half the distance (Dawson et al., 2016). A second measure, Procrustes analysis, applies translation, rotation, and scaling to a "resting" contour to minimize the sum of squared differences between each target contour and the resting state (Goodall, 1991). A third measure, the number of inflection points (NINFL, Preston, McCabe, Tiede, & Whalen, 2019) represents the number of sign changes of a given lingual contour. All three of these metrics have been used to represent curvature across target phonemes while controlling for differences in vocal tract size. Thus, these three metrics will be considered in detail as candidate approaches to measuring tongue complexity in the current dissertation.

Beyond perceptual ratings

We will measure degree of tongue complexity of contours extracted from ultrasound images to determine whether differential patterns can be found between individuals varying in age and clinical presentation. Of key importance, lingual differentiation does not always provide the same information as perceptual ratings of accuracy. Some perceptually-accurate rhotic productions from TD children have been shown to correspond with undifferentiated gestures (Klein et al., 2013). This observation dovetails with former reports of covert articulatory contrasts, a phenomenon in which distinct tongue shape patterns are observed for

targets that are perceptually neutralized (Gibbon, 1999; Gibbon, Dent, & Hardcastle, 1993; McAllister Byun, Buchwald, & Mizoguchi, 2016). In such reports, velar targets were perceived to be neutralized with alveolar targets in child speech, but distinct articulatory configurations were detected via EPG or ultrasound. Conversely, previous reports of covert articulatory errors for stop consonants (Cleland, Scobbie, Heyde, Roxburgh, & Wrench, 2017) and for rhotic targets (Klein et al., 2013) have revealed atypical lingual configurations that are not detected perceptually. These findings highlight how ultrasound may reveal fine details about speech production that are distinct from perceptual-acoustic measures and may therefore provide an additional source of insight for diagnosis and treatment planning.

Measuring somatosensory acuity

Arriving at a thorough understanding of sensorimotor control in child speech also requires further understanding of somatosensory acuity. A few studies have explored the relationship between speech outcomes and oral somatosensory skill, which is typically determined by assessing an individual's ability to use their articulators to detect pressure or vibration or to identify details about an object (Attanasio, 1987). Adolescents with SSD were found to exhibit reduced somatosensory acuity relative to TD peers, as shown with an oral stereognosis task in which an individual identifies the form of an object presented in their oral cavity (Fucci & Robertson, 1971). Similarly, TD adolescents and adolescents with /s/ misarticulation showed greater somatosensory acuity than those who misarticulated rhotics, as shown with an oral form discrimination task (McNutt,

1977). Interpreting these findings within a DIVA framework, performance on oral stereognosis tasks may reflect the degree of specificity of an individual's somatosensory targets and access to somatosensory feedback (Ghosh et al., 2010). Despite the promising set of findings from the 1970s, the topic was less studied in recent years until Ghosh et al. (2010) used an oral somatosensory task involving grating-orientation judgment to look for differences in sibilant production. They found that adults with high somatosensory acuity produced larger /s/~/ʃ/ phonetic contrasts than those with low somatosensory acuity. Considering the lack of recent research for other populations and targets, there is a pressing need for the modern literature to include further investigation of the relationship between somatosensory acuity and speech outcomes.

Understanding the connection between somatosensory capacity and speech outcomes has the potential to offer theoretically-motivated insight into sensory influences on speech production. As both somatosensory acuity and tongue complexity are believed to predict speech outcomes, the two measures may also be correlated with one another. Investigation into the interrelationships among speech outcomes, tongue complexity, and somatosensory acuity (while controlling for auditory acuity) could reveal how each distinct sensorimotor skill individually contributes to speech production.

Aims of dissertation

The overall goal of this research is to determine the extent to which sensorimotor factors predict speech outcomes in children. The first aim focuses on differences in ultrasound-based tongue complexity measures between groups

known to differ in speech production abilities, including young TD children, young children with SSD, and older children with RSE. The second aim examines the relationship between somatosensory acuity and tongue complexity in older children with RSE. The following paragraphs correspond to how each of the three manuscripts addresses the two aims of the current dissertation.

Before it was possible to ask the theoretically-motivated questions about lingual control that form the core of this research, it was necessary to determine what metric(s) of tongue complexity are most suitable for application with child participants. Therefore, the first manuscript lays the foundation for the dissertation in that it justifies the selection of MCI and NINFL as the two metrics of tongue complexity that can most fruitfully be applied to child speech data. First, this study explores phoneme-specific patterns of tongue complexity in an existing sample of adult speech data as a basis for evaluating the extent to which measured values align with expected complexity categories. The study then addresses the question of whether the same patterns are present in a data set of young TD children. This second line of inquiry aligns with the first aim of this dissertation by including a qualitative comparison of tongue complexity in adults versus young TD children.

The second manuscript applies the two identified metrics of tongue complexity (MCI and NINFL) to younger children with and without SSD. This research also addresses the first aim, which is to establish the relationship between tongue complexity and speech outcomes. The analyses in this study explore whether tongue complexity for a variety of phonemes differs over the

course of development in young children, between young children with and without SSD, and between perceptually correct versus incorrect tokens of later-developing sounds. To address whether there are differences in tongue complexity based on perceived accuracy, tongue complexity is compared between correct versus incorrect productions of only the late-developing phonemes /l/ and /ɹ/ in both diagnostic groups.

The third manuscript applies the same two metrics of tongue complexity (MCI and NINFL) to a sample of older children with persistent /ɹ/ misarticulation. This manuscript addresses the first aim in that it explores whether there are differences in tongue complexity based on acoustically measured accuracy of the /ɹ/ sound as produced by children with RSE affecting /ɹ/. Additionally, this study considers treatment effects by including child productions from before and after an ultrasound biofeedback treatment package, which could shed light on motor-based gains that are achieved by individuals undergoing treatment for /ɹ/. This manuscript also addresses the second aim, which is to understand the relationship between somatosensory acuity and tongue complexity. To achieve this aim, we evaluate whether somatosensory acuity is associated with tongue complexity across the same individuals undergoing treatment for /ɹ/ misarticulation while controlling for the better-studied covariate of auditory acuity.

Clinical impact

In summary, this dissertation examines two understudied sensorimotor measures, tongue complexity and somatosensory acuity, in connection with the acquisition of adultlike speech. The two aims of this research comprise critical

steps toward an important goal in the clinical assessment of SSD, namely, to differentiate children with SSD and aberrant sensorimotor skills from those who have simply habituated an incorrect pattern. We posit that degree of tongue complexity as measured from ultrasound images may serve as a novel index of motor execution skill that could help identify those children with SSD and delayed motor development. Likewise, the establishment of a valid approach for measuring somatosensory acuity in young children could help identify those children with reduced ability to access and respond to feedback in the somatosensory domain.

In the long term, this dissertation is expected to form the basis for a program of basic scientific and translational research that could help clinicians gain an improved understanding of the specific sensorimotor factors associated with different speech outcomes in their clients. This information will in turn guide them in treating children with various clinical profiles of SSD. Specifically, children with delayed motor execution skill may be best suited for a traditional treatment approach that incorporates the principles of motor learning (Maas et al., 2008), whereas those with typical motor development may be best suited for a phonetic-perceptual approach to treatment (Gibbon, 1999). Similarly, children with reduced somatosensory acuity could be assigned to a form a treatment that enhances that specific deficit area (e.g., ultrasound biofeedback to draw direct attention to placement of the articulators). In this way, the present dissertation and proposed follow-up research could enable a shift toward greater consideration of individual sensorimotor factors in order to develop personalized learning

approaches (e.g., Wong, Vuong, & Liu, 2017) for children with SSD. Tailoring intervention in this way could result in faster responses to treatment, providing benefit to both clinicians and children with SSD.

MANUSCRIPT 1: Comparing metrics for quantifying children's tongue shape complexity using ultrasound imaging

Abstract

Speech sound disorders can pose a challenge to communication in children that may persist into adulthood. As some speech sounds are known to require differential control of anterior versus posterior regions of the tongue body, valid measurement of the degree of differentiation of a given tongue shape has the potential to shed light on development of motor skill in typical and disordered speakers. The current study sought to compare the success of multiple techniques in quantifying tongue shape complexity as an index of degree of lingual differentiation in child and adult speakers. Using a pre-existing data set of ultrasound images of tongue shapes from adult speakers producing a variety of phonemes, we compared the extent to which three metrics of tongue shape complexity differed across phonemes/phoneme classes that were expected to differ in articulatory complexity. We then repeated this process with ultrasound tongue shapes produced by a sample of young children. The results of these comparisons suggested that a modified curvature index and a metric representing the number of inflection points best reflected small changes in tongue shapes across individuals differing in vocal tract size. Ultimately, these metrics have the potential to reveal delays in motor skill in young children, which could inform assessment procedures and treatment decisions for children with speech delays and disorders.

Introduction

In the speech of typically developing children (Fletcher, 1989) and children with speech sound disorders (Gibbon, 1999), limitations in the child's capacity for isolated control of anterior versus posterior lingual regions may play an important role in nonadultlike speech patterns. This capacity for different regions of the tongue to operate semi-independently is referred to as "lingual differentiation"; gestures that lack a typical degree of independent lingual control may be described as "undifferentiated gestures" (Gibbon, 1999). In many cases, undifferentiated tongue shapes are associated with perceptually incorrect productions (i.e., a substitution or distortion) in child speech. However, for both typically developing (TD) children and children with speech sound disorder (SSD), it is possible for degree of lingual differentiation to dissociate from perceived accuracy for a given production (Gibbon, 1999). Most notably with children with SSD, previous literature has described cases in which undifferentiated or atypical gestures are present in productions that are perceptually transcribed or rated as accurate, sometimes termed "covert error" (Cleland et al., 2017). For other individuals, there is documented physiological evidence of "covert contrast", such as when perceptually neutralized productions are produced with measurably different tongue shapes (Gibbon, 1999). For example, for a four year-old child with SSD who exhibited the phonological pattern of alveolar backing, Gibbon (1999) found that /g/ productions were produced with appropriate differentiated velar contact with the palate, whereas /d/ productions were produced with an undifferentiated shape involving both velar and alveolar contact. While the covert patterns described here refer to populations

with SSD, children may also exhibit covert contrast over the course of typical phonological development, as originally described with respect to voicing contrasts by Macken and Barton (1980). Therefore, assessing a child's degree of lingual differentiation has the potential to provide information about motor maturation that cannot be obtained from transcribed speech alone.

The goal of the present study is to lay a foundation for research to quantify TD children's degree of lingual differentiation, which we operationalize in the present study as "degree of tongue shape complexity." Previous attempts to automate quantification of the degree of tongue shape complexity using tongue contours from ultrasound imaging have yielded promising results. Dawson et al. (2016) used multiple methods for quantifying degree of tongue shape complexity in adult speakers producing various phonemes. Preston et al. (2019) applied an additional ultrasound-based metric of tongue shape complexity to child speakers. In the present study, we considered three of these established approaches to quantification of tongue shape complexity and applied them to adult data representing a range of phonemes. We first evaluated the extent to which the three measures differentiated between phonemes/phoneme classes theoretically expected to differ in articulatory complexity. Then we examined whether the same patterns found with adults were also present when the three measures were applied to ultrasound tongue shape data from children, whose articulation is known to be more variable than that of adults (Goffman & Smith, 1999).

Expected complexity of phoneme classes

Before applying metrics of tongue shape complexity to child or adult speech, it is essential to consider what degree of tongue shape complexity might be expected for a given phoneme. Not all speech sounds are expected to be produced with complex tongue shapes; for instance, it is widely agreed that vowels are articulatorily simple and the liquid sounds /l/ and /ɹ/ are articulatorily complex (Kent, 1992). In the context of a project to develop measures of tongue shape complexity, it is difficult to avoid circularity when defining the expected complexity of a phoneme. One alternative is to draw on taxonomies that describe the order in which phonemes are acquired in typical child speech development, since later-developing sounds tend to involve more complex tongue shapes. Such taxonomies are commonly derived from transcription-based studies of the perceived accuracy of children's speech at different ages. Clinicians assessing English-speaking children for speech disorders commonly make reference to Shriberg's (1993) system that groups consonants into early, middle, and late stages based on data from 64 children ages 3-6 with speech delays. The "early eight" include /m/, /n/, /p/, /b/, /d/, /w/, /j/, and /h/; the "middle eight" include /ŋ/, /t/, /k/, /g/, /f/, /v/, /ʧ/, /ʤ/; and the "late eight" include /θ/, /ð/, /s/, /z/, /ʃ/, /ʒ/, /l/, and /ɹ/. More recently, Crowe and McLeod (2020) conducted a systematic review of fifteen studies comprising over 18,000 children acquiring American English and reported a slightly different set of three stages: the "early 13" include plosives, nasals, and glides, the "middle 7" include affricates, unvoiced fricatives, and laterals, and the "late 4" include rhotics and voiced fricatives.

Other developmental taxonomies have been established with more explicit reference to the articulatory complexity of speech sounds. It is important to acknowledge at the outset that not all articulatory complexity is tongue shape complexity; some sounds have increased complexity because they require coordination of oral articulatory gestures with glottal gestures or opening/closing of the velopharyngeal port. We begin with a broad view that encompasses all aspects of articulatory complexity, and subsequently narrow our focus to the specific topic of tongue shape complexity. Studdert-Kennedy and Goldstein (2003) described a hierarchical relationship among classes of phonemes that corresponds with how much coordination among articulatory gestures is needed to achieve accurate production. Young children were reported to first show mastery of voiceless stops, nasals, glides, and /h/, and then later added voicing contrasts as they developed the ability to coordinate laryngeal gestures with lingual gestures. The third stage occurred when children were able to coordinate jaw height and/or degree of constriction with lingual gestures in order to produce fricatives and affricates. They described the final stage as occurring when multiple lingual constrictions began to occur, allowing for the production of liquid consonants including /l/ and /ɹ/ (Studdert-Kennedy & Goldstein, 2003).

In a similar description of stages of speech development in English, Kent (1992) described speech sounds according to the shape of the tongue and the type of movement involved in production. For consonants, Kent (1992) presented four stages of consonant sound development, classified according to the degree of 'ballistic' versus 'controlled' movement involved in articulation. He described

how young children in the first stage produce consonants with mostly 'ballistic' movement, characterized by short durations with high-velocity accelerations and decelerations, as in /p/, /m/, and /n/. Some consonants in this stage also feature "ramp" movements, which involve slow movements of relatively stable velocity and long duration, as in /w/ and /h/. The next class of sounds to develop involves more rapid ballistic movements (/b/, /k/, /g/, /d/) and ramp movements (/j/), and the emergence of the fricative /f/. In the third stage, an additional rapid ballistic sound (/t/) is added, along with controlled movements that allow voicing distinctions as well as the complex tongue shapes associated with /ɹ/ and /l/. Kent (1992) described the fourth and final stage as comprising the additional fricative sounds that require precise focal control at the point of lingual constriction.

Although there are differences across these four hierarchies of consonant development based on perceived accuracy and/or articulatory development (see *Table 2.1*), there is general agreement that nasals, voiceless stops, glides, and /h/ are developed relatively early due to their reduced level of articulatory complexity. (Vowels are generally omitted from such taxonomies on the assumption that they are even simpler than these early consonants, and thus earlier-developing.) It can also be observed that consonants produced using a labial place of articulation tend to develop before the same classes of sounds produced with a lingual constriction. Finally, there is general agreement that consonants requiring a single lingual constriction, such as /t/ and /k/, are usually acquired before sibilants produced with lingual grooving, such as /s/ and /z/, and also before liquids that require multiple lingual constrictions, such as /l/ and /ɹ/.

In their study specifically focusing on tongue shape complexity, Dawson et al. (2016) established multiple methods for quantifying tongue shape complexity and compared the metrics' collective ability to classify tongue shapes into complexity classes. Unlike the other proposed hierarchies, the complexity classes from Dawson et al. (2016) were not intended to represent a developmental hierarchy. They included vowels and did not consider consonantal voicing contrasts (i.e., with the exception of /θ/, only voiced consonants were included in their classification system). Their *low complexity* category included all unrounded vowels with a single lingual constriction, including /ɑ/, /æ/, /ɪ/, /ʌ/ and /ɛ/. Their *medium complexity* group included sounds involving lip rounding, including /w/ and /u/, sounds with lateral bracing, including /j/, and sounds with a constriction formed with the tongue dorsum, including /g/. Their *high complexity* group included all sounds with a constriction of the tongue tip, including /d/ and /l/, sounds with more than one lingual constriction, including /ɹ/, and all fricatives, including /z/, /θ/, and /ʒ/. Dawson et al. (2016) provided empirical evidence from ultrasound-based midsagittal lingual contours in support of these *a priori* categories.

The categories proposed by Dawson et al. (2016) agree in several respects with the other four phoneme hierarchies (see *Table 2.1* for a summary), including classifying fricatives, laterals, and rhotics as having relatively higher complexity than other consonants, including nasals, glides, and stops. The most notable discrepancy is that the alveolar stops /t/ and /d/ are placed in the early or middle groups by Shriberg (1993), Crowe and McLeod (2020), and Studdert-Kennedy

Table 2.1. Hierarchies of consonant complexity.

Shriberg (1993)	Crowe & McLeod (2019)	Studdert-Kennedy & Goldstein (2003)	Kent (1992)	Dawson, Tiede, & Whalen (2016)
"early" /m/, /n/, /p/, /b/, /d/, /w/, /j/, /h/	"early" /b/, /n/, /m/, /p/, /h/, /w/, /d/, /g/, /k/, /f/, /t/, /ŋ/, /j/	voiceless stops, nasals, glides, /h/	/p/, /m/, /n/, /w/ /h/	"medium complexity" /w/, /j/, /g/
"middle" /ŋ/, /t/, /k/, /g/, /f/, /v/, /ʧ/, /ʤ/	"middle" /v/, /ʤ/, /s/, /ʧ/, /l/, /ʃ/, /z/	voicing contrasts	/b/, /k/, /g/, /d/, /j/, /f/	
"late" /θ/, /ð/, /s/, /z/, /ʃ/, /ʒ/, /l/, /ɹ/	"late" /ɹ/, /ð/, /ʒ/, /θ/	fricatives, affricates	/t/, voicing distinctions, /ɹ/, /l/	"high complexity" /d/, /l/, /ɹ/, /z/, /θ/, /ʒ/
		/l/, /ɹ/	additional fricative sounds	

and Goldstein (2003). However, /t/ is developed relatively late according to Kent (1992) and /d/ is included in the "high complexity" category according to Dawson et al. (2016). The present research drew on the categories from Dawson et al. (henceforth, "Dawson categories") because they broadly align with existing hierarchies of articulatory development and because we re-analyzed data from their original study. However, we kept the other taxonomies in mind throughout our analyses, and we paid particular attention to the patterning of alveolar stops due to the discrepancy in their characterization across previous studies.

Instrumental measures of motor control

The preceding discussion suggests that tongue complexity could be a valuable measure for assessing motor skill in children with a suspected delay or disorder in speech development. It is thus important to consider different approaches that can be used to measure tongue complexity or other indicators of

articulatory skill. Gibbon's (1999) foundational work on lingual differentiation used electropalatography (EPG), which is useful because it makes readily visible what areas of palatal contact are present, allowing researchers to infer what lingual regions are being used to form a constriction. EPG was instrumental in the discovery of developmental decreases in the amount of broad linguopalatal contact for a variety of targets (Fletcher, 1989). Gibbon (1999) distinguished between stop consonants produced in a "differentiated" fashion (i.e., with contact isolated to a sub-region of the palate) and those produced in an "undifferentiated" fashion (i.e., with broad palatal contact), and found covert contrasts between stops produced with velar versus simultaneous alveolar and velar places of articulation. EPG has also been instrumental in revealing linguopalatal contact patterns in lateral bracing in sibilants (Gibbon, Hardcastle, & Dent, 1995). Although there is evidence supporting the use of EPG for diagnosis and treatment of speech disorders (Bernhardt, Bacsfalvi, et al., 2005), clinical application of the method is limited by the high cost and time delay required to manufacture individually customized palatal prostheses. Additionally, the approach is not considered well-suited for children with growing vocal tracts, who may require several palates over the course of development (Gibbon, 1999). Therefore, there is a need for alternative approaches to measuring lingual differentiation that are more accessible and better suited for young children.

Ultrasound imaging is an increasingly available and affordable alternative that is also minimally invasive and thus suitable for use with small children. Data collected from EPG and ultrasound are not directly comparable because EPG

provides discrete information about palatal contact patterns, whereas ultrasound provides a continuous representation of tongue shape in one anatomical plane (e.g., midsagittal). Despite this fundamental difference, it is reasonable to suggest that insights from the EPG-based lingual differentiation literature may dovetail nicely with the current efforts to quantify tongue shape complexity using ultrasound. Namely, ultrasound has been used to reveal covert contrasts for velar targets (McAllister Byun et al., 2016) as well as covert errors in perceptually accurate /k/ and /t/ productions (Cleland et al., 2017). It is highly probable that these ultrasound-based findings provide insight into the same covert articulatory phenomena previously observed using EPG. These findings suggest that both EPG and ultrasound can provide articulatory information that is distinct from and supplementary to readily available ratings of perceived accuracy. Extending ultrasound imaging of the tongue into the clinical domain, differences in degree of lingual differentiation have been quantified using ratio-based measures (Klein et al., 2013; Ménard, Aubin, Thibeault, & Richard, 2012; Zharkova, Gibbon, & Hardcastle, 2015). While such approaches are helpful for describing shape and position of contours with one lingual constriction, they are not suitable for quantifying differences among contours with multiple lingual constrictions, such as /l/ and /ɹ/. Instead, qualitative descriptions have been used to describe differences in tongue shape between individual children with and without SSD producing rhotics (Klein et al., 2013). In light of its relative accessibility, its potential to provide insight into a child's stage of motor development, its suitability for measuring a variety of speech targets, and its clinical applications,

the ultimate goal of the present study is to use ultrasound to quantify tongue shape complexity in young children.

Although in the present study we chose to focus on measuring tongue shape complexity, it is important to note that previous research has identified other means of quantifying motor development, including measures of articulatory coupling and movement variability using such methods as electro-magnetic articulography (EMA). Tracking kinematic measures of lip and jaw movement, Green, Moore, Higashikawa, and Steeve (2000) found incremental increases in temporal coupling of these two articulators with increasing age. Children also exhibit a relatively high degree of motor variability that decreases over the course of development as lip and jaw motor targets are refined (Goffman & Smith, 1999; Grigos, 2009). Similarly, children with motor SSD show greater lip, jaw, and tongue tip movement variability than TD children (Terband, Maassen, Van Lieshout, & Nijland, 2011). However, because EMA is limited to the anterior vocal tract, it is not appropriate for tracking posterior lingual constrictions, and therefore is not optimal for questions about late-emerging, complex sounds like the liquids /l/ and /ɹ/. It also is more invasive and therefore more challenging to use with young children than ultrasound.

Therefore, we arrive at the present need for a valid and accessible metric that represents the degree of differentiation of lingual contours, including contours with multiple constrictions. As highlighted in the preceding discussion of taxonomies of articulatory development, it is important to keep in mind that the expected degree of lingual differentiation differs across phonemes; for instance,

most vowels are produced with simple tongue shapes. Since the term "undifferentiated" implies an absence of differentiation that ought to be present, we favor the term "tongue shape complexity" (as used by Dawson et al. [2016]) because it can neutrally characterize both within- and across-speaker differences without suggesting an expected tongue shape. In addition, continuous measures of tongue shape complexity might be more appropriate than a binary differentiated/undifferentiated distinction to evaluate small articulatory differences across phonemes and groups of speakers.

Ultrasound measurement of tongue shape complexity

Ultrasound is used to visualize boundaries between tissues with different densities, such as the boundary between the surface of the tongue and the air in the oral cavity, as detailed in Stone (2005). An ultrasound transducer, or probe, is placed beneath the chin and piezoelectric crystals inside the probe emit high-frequency sound waves in a fan-like shape through a section of the tongue (either sagittal or coronal, depending on the orientation of the probe). The time that it takes sound to return to the probe after being reflected by the density boundary is used to generate an image of the surface of the tongue, which appears as a white line. Sound waves do not readily pass through bone, which therefore appears as a black shadow in ultrasound images. When imaging in a midsagittal section, it is desirable for the field of view to extend from the shadow of the hyoid bone in the posterior aspect to the shadow of the mandibular bone anteriorly.

A variety of approaches have been used to quantify the shape of a lingual contour that has one major place of constriction. The simplest strategy is to derive

measures from raw coordinates representing the surface of the tongue. However, these coordinates are relative to the position of the ultrasound probe, resulting in noise in the signal if the probe moves independently from the head. In order to compare any two sets of raw coordinates within or between individuals, it is essential to control for head movement or otherwise determine that the two contours have identical head position and probe orientation. One approach is for the probe to be physically stabilized relative to the head, as with headsets and collars (Cleland et al., 2017; Derrick, Carignan, Chen, Shujau, & Best, 2018; Stone, 2005), but this equipment can be heavy and uncomfortable and is not well tolerated by all participants, especially children. An alternative approach is to allow the head to move freely, but to measure its position and orientation relative to the probe using tracked visual or infrared sensors (Kabakoff, Harel, Tiede, Whalen, & McAllister, 2021; Whalen et al., 2005). Such measurements can then be used to normalize observed lingual position to a consistent head-centric coordinate system, or alternatively used to identify productions that are misaligned and should therefore be removed from the data set.

Instead of or in addition to using preventative or corrective approaches to control for head movement, it is possible to derive measures that are relatively robust to the rotation and translation introduced by head movement. Previous research has proposed two such measures, *curvature degree* and *curvature position,* derived from the length of the contour from the mandible to the hyoid shadow and the height of the contour at the highest vertical point perpendicular to the length (Ménard et al., 2012; Zharkova et al., 2015). As these measures are

ratios of two segment lengths, they intrinsically normalize across speakers of different sizes. While these metrics are equipped for describing one lingual constriction, as in most vowels and stop consonants, they cannot distinguish contours with multiple lingual constrictions, such as /l/ and /ɹ/. This highlights the need for metrics that can capture the degree of curvature along multiple lingual constrictions while remaining robust across individuals differing in vocal tract size.

Dawson et al. (2016) developed and compared various approaches to obtaining continuous metrics that normalize for differences in vocal tract size and head movement with the goal of quantifying the complexity of lingual contours across a variety of phoneme targets in adults. They sought to determine which metrics best classified adult productions into the pre-established low, medium, and high complexity categories described above. The primary metrics of tongue shape complexity were a modified curvature index (MCI) and a Procrustes analysis. MCI is the average of unsigned curvature integrated at each point along the length of the arc of a traced lingual contour (Dawson et al., 2016); it differs from another published curvature index (Stolar & Gick, 2013) in that the curve parameterization is used as the reference rather than the x-axis (which would require head stabilization for the values to be interpretable). MCI for a given tongue contour is determined by first computing the absolute curvature (the reciprocal of the tangent circle) at each of the normalized equidistant points along an outline of the tongue surface, and then integrating across these. The Procrustes analysis utilizes a lingual contour at rest to obtain a baseline measure intended to

represent a minimal degree of tongue shape complexity (Dawson et al., 2016). Dawson et al. (2016) described the "resting" contour as a pre-phonatory position in which the tongue "lay flat in the mouth, with no palate contact." Tongue contours obtained during target phonemes are superimposed over this resting contour, and then translation, rotation, and scaling are applied to minimize the sum of squared differences between each frame and the resting state (Goodall, 1991). Finally, Dawson et al. (2016) also considered an analytical approach in which a Discrete Fourier Transform (DFT) was used to transform the tangent angles of a tongue contour into a characterization of tongue shape as a sum of its spatial frequency components (Liljencrants, 1971). DFT yields coefficients with real and imaginary components, of which the first corresponds to a wavelength equal to the contour length and higher coefficients to multiples of this frequency. The real component of each coefficient provides phase information (at what point along the vocal tract a constriction occurs); the imaginary component provides the magnitude of the constriction. In Dawson et al. (2016), including all three coefficients (C1, C2, C3) did not improve categorization. DFT did provide more consistent categorizations than MCI and Procrustes, but did not make for an 'intuitive' interpretation in terms of complexity.

Dawson et al. (2016) used linear discriminant analysis to determine which metrics or combination of metrics best classified various phonemes into complexity classes. In their analysis, Dawson et al. (2016) found that the metrics that were best at independently grouping individual productions into their proposed complexity categories were the imaginary component of C1 (77%

accuracy), Procrustes (62% accuracy), and MCI (56% accuracy). However, they determined that the combination of the real and imaginary components of C1 from the DFT together was an even better classifier (81% accuracy), and that adding MCI and Procrustes to this combination improved classification accuracy even more (83% accuracy). However, it is important to note that although the imaginary component of C1 was successful at classifying tongue shapes into complexity categories in Dawson et al. (2016), it is difficult to interpret and compare DFT coefficients across speakers in the absence of how a given tongue shape maps into idiosyncratic vocal tract morphology, information not available from ultrasound alone. Based on these considerations, only MCI and Procrustes were examined as candidate measures of tongue shape complexity in the present analyses.

Preston et al. (2019) proposed an additional metric for quantification of tongue shape complexity that is robust across differences in vocal tract size: an ordinal measure that represents the discrete number of inflection points (NINFL) determined by the number of curvature sign changes of a given lingual contour. To avoid including inflections due to small local changes in curvature, only changes exceeding a consistent threshold are counted. Comparing NINFL values of /ɹ/ sounds produced by school-aged children with and without SSD, Preston et al. (2019) found that children with /ɹ/ misarticulation had lower NINFL values than TD children. Additionally, NINFL values correlated with /ɹ/ accuracy ratings, such that higher values were associated with higher perceived accuracy. Finally, for those children enrolled in treatment for /ɹ/ misarticulation, lingual

contours showed higher NINFL values after treatment than before treatment. Although Preston et al. (2019) did not apply NINFL to other sounds, its success in quantifying changes in /ɹ/ production suggests that it could be useful to distinguish among phonemes involving dual lingual constrictions, such as /l/, and phonemes involving less complex tongue shape.

The current study

The present study first compared the three above-mentioned metrics in the sample of adult speakers producing a variety of targets in Dawson et al. (2016), then applied them to child data with the ultimate goal of identifying the metrics that best represent degree of tongue shape complexity in children. The specific objectives of the current study are as follows:

1. To determine the extent to which the three metrics distinguish various adult speech targets expected to differ in articulatory complexity based on established taxonomies.
2. To determine the extent to which patterns of tongue shape complexity found with adults were also present in children for whom relatively late-developing (and therefore lingually-complex) targets may still be emerging.

For the first objective, we applied the three metrics to the adult participants from Dawson et al. (2016) to see how well they separated the adult productions into phonemes and phoneme classes. The rationale for conducting this initial re-analysis of the adult data was threefold. First, this analysis was intended to draw attention to any metric-specific categorization patterns for phonemes and

phoneme classes, which were not readily apparent in Dawson et al. (2016) due to their focus on overall classification accuracy for the metrics. Second, the present analysis included NINFL, which was not one of the metrics considered in Dawson et al. (2016). Third, adult productions are known to show more articulatory stability than child configurations (Goffman & Smith, 1999), so it follows that analysis of tongue shape complexity in children would be premature without detailed knowledge of what patterns ought to be present in mature adult speech. For the second objective, we applied the same three metrics to a new data set of young TD children to determine whether the same patterns related to phoneme and phoneme class found in adults were also present in children.

We predicted that for both adults and children, we would see general agreement between measures of tongue shape complexity and articulation-based schemes of phoneme acquisition, such that later-emerging phonemes would be associated with higher tongue shape complexity (i.e. tongue complexity would be lowest for vowels, higher for glides, and highest for liquids). For children for whom late-developing targets may still be emerging, we anticipated a reduced degree of separation across phonemes due to articulatory simplifications that may especially affect late-developing targets.

Identifying a measure that agrees with existing schemes for classifying articulatory complexity will increase confidence in the clinical utility of these measures. If our proof-of-concept analyses suggest that these measures may also be valid for child data, this would support further research using the selected metrics to quantify differences within and between child speakers. Ultimately,

such measures could support clinical efforts to identify the relative contribution of motor skill to a child's error patterns, with implications for diagnosis and treatment planning.

Methods

Participants

Adult data set

The adult data set from Dawson et al. (2016) was used with permission. This data set included 1125 productions from six participants between the ages of 24 and 45 with no history of speech or language impairment who were seen at the Graduate Center of the City University of New York (CUNY). The target vowels /ɑ/, /æ/, /ɛ/, /ɪ/, /u/, and /ʌ/ were produced in a /bVb/ context, and the target consonants /j/, /w/, /ɡ/, /d/, /z/, /ʒ/, /l/, /ɹ/, and /θ/ were produced in an /ɑCɑ/ context. After rehearsal of the complete set of stimuli, participants produced two sets of at least six repetitions of each stimulus, elicited in a random order. Ultrasound recordings were collected with an Ultrasonix SonixTouch with a C9-5/10 microconvex transducer (frequency range 5-9MHz, 10 mm footprint) at 60 frames per second. A heavy-duty metal stand with a spring-loaded probe arm was used to minimize probe movement relative to each participant's jaw. The frame selected for measurement was the frame closest to the acoustic midpoint for vowels and the point of maximal constriction (i.e., greatest lingual displacement) for consonants. See *Table* 2.2 for details about the adult participants after exclusions (as described below), including gender and total number of tokens in the final data set.

Child data set

The child data set included 1132 productions from 17 typically developing children who participated in an evaluation at one of three sites, including Molloy College, Haskins Laboratories (Yale University), and Syracuse University. Data collection from this study was carried out in accordance with the Molloy College Institutional Review Board (no protocol ID provided), the Yale University Human Research Protection Program Institutional Review Boards (protocol ID #1610018484), and the Syracuse University Institutional Review Board for the Protection of Human Subjects (protocol ID #16-282). The participating children had a mean age of 5;2 (range 4;2-6;3) and included 11 females and 6 males. All children had normal hearing and no history of speech or language impairment. However, many of these children produced at least some errors on the late-developing sounds /ɹ/ and /l/ (as described below), as these phonemes were still emerging along a developmentally appropriate trajectory.

At Molloy College, ultrasound recordings were collected with a Siemens Acuson X300 with a C8-5 wideband curved array transducer (frequency range 3.1–8.8 MHz, 25.6 mm footprint, 109 degree field of view) at 43-49 frames per second with 60-70 mm depth. At Haskins, a Siemens Acuson X300 was used with a C6-2 wideband curved array transducer (frequency range 1.8–6 MHz, 73.0 mm footprint, 90 degree field of view) at 36-37 frames per second with 80 mm depth. At Syracuse University, a Telemed Echoblaster 128 was used with a PV 6.5 wideband curved array transducer (frequency range 5–8 MHz, 156 degree field of

view) at 21-25 frames per second with 110 mm depth).[1] Ultrasound video recordings were captured at 60 frames per second on a PC through an AverMedia video capture card at Molloy and Haskins, and at 35 frames per second with Debut (NCH Software) at Syracuse. The ultrasound probe was placed in a microphone stand while the clinician supported alignment of the probe with each child's head. In addition, blue dots were placed on the forehead, nose, lips, chin, and on the ultrasound probe in order to measure the alignment of the probe with the head (see ultrasound measurement section below). Children were initially familiarized with the pictures used for elicitation prior to placement of the ultrasound probe, with the evaluating clinician providing cues or modeling the word as needed until the child could name each image. The researcher monitored the ultrasound image during data collection, and if they had concerns about the quality of the ultrasound image during a given production (e.g., due to movement of the child's head relative to the probe), an additional production was prompted. Sixteen words were elicited three times each in random order for a total of forty-eight productions.[2] Consonants were targeted in initial position and included /j/ in

[1] The divergent frame rates used with the different systems had the potential to affect our ability to identify the optimal frame within a given acoustic interval. At our lowest frame rate of 21 frames per second, the selected frame could be at most 48 ms from the true frame of interest, which is judged to be sufficient for the present non-dynamic analysis. Although reduced zoom depth may result in fewer pixels available, MCI and NINFL computations are made from overlayed anchors and do not depend on the available number of pixels.

[2] Because re-elicitations were possible with the young child sample, more than 48 productions were elicited from some child participants. However, some elicitations were later determined to have misaligned ultrasound images or were otherwise unclear (as indicated in the ultrasound measurement section) and were removed from the final data set.

"yam", /w/ in "wake" and "wing", /k/ in "cape", "cat", "coat", "key", /t/ in "tape", "tea", and "toe", /l/ in "lake" and "lamb", and /ɹ/ in "rake", "rat", "ring" and "rope." As word-initial consonants were the present focus, final consonants were not analyzed. However, we did include the two monophthongs (/æ/ and /ɪ/) in words with final consonants in the present analysis to allow for a more complete extension of Dawson et al. (2016). As such, the target vowels included /æ/ in "cat", "lamb", "rat", and "yam" and /ɪ/ in "ring" and "wing." See *Table* 2.2 for details about the child participants, including age, gender, site, and total number of tokens in the final data set. Although accuracy ratings were not performed as part of the present study, accuracy ratings based on narrow transcription (in which distortions were classified as errors) from Kabakoff et al. (2021) indicated that these same typically developing children produced /æ/ with 98.1% accuracy, /ɪ/ with 86.4% accuracy, /j/ with 92.3% accuracy, /w/ with 100% accuracy, /k/ with 98.6% accuracy, /t/ with 98.4% accuracy, /l/ with 84.4% accuracy, and /ɹ/ with 42.3% accuracy. See Kabakoff et al. (2021) for more information on error patterns for these individuals.

Table 2.2. Participant details

Participant	Age	Gender	Site	Tokens
A01M	Adult	M	CUNY	156
A02F	Adult	F	CUNY	202
A03F	Adult	F	CUNY	184
A05M	Adult	M	CUNY	193
A06M	Adult	M	CUNY	179
A07M	Adult	M	CUNY	194
01F	5;6	F	Haskins	66
02F	5;11	F	Haskins	21
03M	5;6	M	Haskins	58
04F	5;6	F	Haskins	46
05M	4;2	M	Haskins	19
06M	4;11	M	Haskins	56
07F	5;11	F	Haskins	70
08F	5;9	F	Haskins	35
10F	4;3	F	Molloy	50
11F	6;0	F	Molloy	45
12F	4;9	F	Molloy	62
13M	6;3	M	Molloy	85
14F	4;3	F	Molloy	34
15F	4;5	F	Molloy	26
16M	5;3	M	Syracuse	38
17F	5;5	F	Syracuse	62
18M	4;6	M	Syracuse	65

Ultrasound measurement

All processing of ultrasound data from adults was performed at CUNY following the protocol described in Dawson et al. (2016). All processing of ultrasound data from children was performed at New York University as part of a larger study that included children with SSD. The procedures used for the child data are described briefly below; see Kabakoff et al. (2021) for additional detail.

Midsagittal ultrasound probe alignment was quantified using a procedure in which blue dots were placed along the vertical midline of the child's face and the probe. The position of the dots was automatically tracked in frontal-view video recorded concurrently with ultrasound data collection. This video was used in a Matlab (MathWorks Inc., 2000) script that temporally aligned the video of

the child's face with the ultrasound video using cross-correlation of their mutual audio. The script then flagged video frames for further inspection if the tracked blue dots indicated more than one standard deviation of displacement across the child sample from Kabakoff et al. (2021). This threshold was 15.4 mm of lateral displacement or 13.3° of angular displacement of the probe relative to the face. Using this method, 24.4% of frames (276/1132) were discarded due to lateral misalignment (170/1132) and/or angular misalignment (197/1132), leaving 856 tokens.

For all sound files, trained university students who had taken courses in linguistics or phonetics and had received project-specific training viewed waveforms and spectrograms in Praat (Boersma & Weenink, 2019) in order to mark the relevant sonorant and obstruent intervals in the time-synced TextGrid file and label them by target phoneme. Each marked acoustic interval in the TextGrid file was viewed in the time-synced ultrasound video in Matlab (MathWorks Inc., 2000) using GetContours (Tiede, 2016), an ultrasound annotation program that supports navigation to the first frame within marked target intervals. The trained university students selected the frame judged to most clearly represent each target phoneme and placed sixteen anchors along the underside of the white line visible on the ultrasound image. The software then automatically redistributed the points evenly along the traced contour, and the evenly distributed points were automatically extrapolated into 100 x- and y-

coordinates. See *Figure 2.1* for a sample set of sixteen anchor points for an /ɹ/ target produced within the word “rope.”

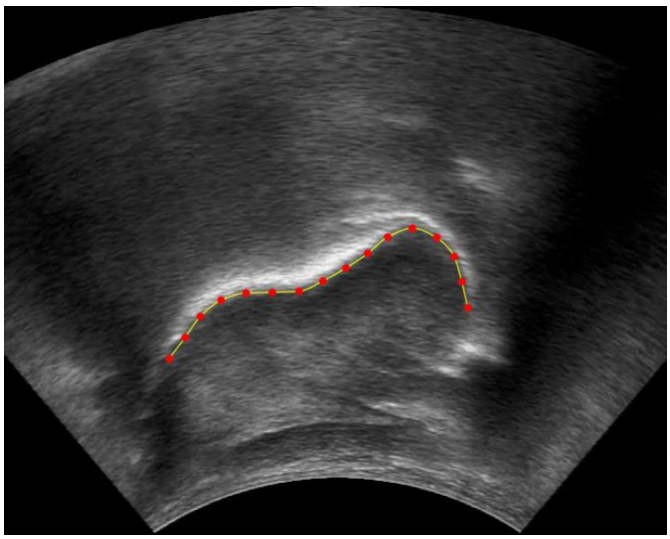

Figure 2.1. Sample set of sixteen evenly distributed anchor points traced in GetContours.

After initial data processing by students, a graduate student with specialized training in phonetic analysis assured consistency across ultrasound files by verifying that all target productions were traced and that all traces reflected the entire visible contour. For most frames, this meant that the tracing of the tongue’s surface should extend from the hyoid shadow to the mandibular shadow. The student specialist discarded any frames that they subjectively judged to be off-center or unclear and retraced any frames that were not traced fully. As such, for the minority of cases where both shadows were not visible, tracings were either judged to represent both posterior and anterior lingual regions fully, or they were discarded. The first author exported all remaining contours (i.e., sets of 100 coordinates), read the coordinates into RStudio (RStudio Team, 2019), scaled (z-score) both the x- and y-axes within speaker and target, and plotted the

coordinates for a final round of quality assurance. Consensus was reached by the first author and the student specialist that 16 tokens should be removed because they exhibited high degrees of perseverative coarticulation, as with 8 /æ/ productions and 7 /ɪ/ productions that showed multiple constrictions following /ɹ/ (in "raft" and "ring"). Additionally, one /k/ exclusion was made (from 15M) because the contour was dissimilar to the shapes of the speakers' other /k/ productions; this was thought to be attributable to a difference in view range where the most posterior regions of the contour was not captured. After this process of removing outliers based on visual inspection, there were a total of 840 tokens in the child data set.

Tongue shape complexity measurements

For both the adult data set and the child data set, MCI and Procrustes metrics were calculated using a custom script (Dawson, 2016) in Python (Python Software Foundation, 2016), and NINFL was calculated using a custom Matlab script (ComputeCurvature). Any NINFL value exceeding five (n = 15 for adults; n = 1 for children) was discarded as unlikely to be a valid representation of a possible tongue shape, following the procedure described in Preston et al. (2019). Similarly, all MCI values exceeding six (n = 5 for adults, n = 1 for children) were removed based on the distribution of MCI values published in Dawson et al. (2016). After exclusion of these outliers, the total number of adult contours was 1108 and the total number of child contours was 838. After calculation of the three metrics, all subsequent analyses were performed using the RStudio interface to R (R Core Team, 2019).

Analyses

Our first objective was to evaluate how well each of the three established metrics of tongue shape complexity, taken individually, classifies tongue contours in the adult data set into phonemes and into pre-established complexity classes. We qualitatively inspected whether distributions of values for each of the selected metrics were distinct across individual phonemes and across natural classes of phonemes. This represents a distinct objective from that of Dawson et al. (2016), which used linear discriminant analysis to find the metric or combination of metrics that yielded the highest classification accuracy, but did not conduct a detailed examination of phoneme-specific patterns with reference to expected articulatory complexity.

Our second objective was to determine whether the patterns found in the adult data set would also be found in children for whom complex targets may still be emerging. As for the adult data, we visually examined whether the distributions of complexity scores for the selected metrics would distinctly categorize the child productions into phonemes and into complexity categories. Since the adult data were elicited in nonwords with a constant phonetic context and the child data were elicited in real words with varying vowels and coda consonants, quantitative comparisons between the two data sets were not possible; therefore, only qualitative comparisons were reported.

Results

For our first objective looking at tongue shape complexity patterns across phonemes in the adult data set, notched boxplots of the three metrics for each

target phoneme were created to allow visual estimation of the extent to which MCI, NINFL, and Procrustes values differed across phonemes. *Figure 2.2* shows complexity values for each target, pooled across participants, and sorted in increasing order by median. Notches represent the confidence interval around the median. While MCI and Procrustes are continuous metrics that can be sorted meaningfully by the median value for each target phoneme, NINFL is an ordinal metric with values ranging from one to five that would lead to many instances of the same whole number when attempting to sort by the median values. Therefore, in order to avoid such ties between NINFL distributions with the same median, it was necessary to ran-order the mean values for each speaker by target phoneme. Calculating the mean in this way leads to fractional values that characterize the NINFL patterns of a given speaker. Therefore, the notched boxplots for NINFL show the median of these means, which characterizes speaker patterns more effectively than if we had plotted the median of the raw ordinal values. Horizontal jitter is added to all three plots. To avoid overlapping values, vertical jitter is added to the ordinal NINFL plot. The boxes are colored by the Dawson categories. As can be seen in the figure, in most cases, the notches around the ranked medians overlap substantially and thus do not fully separate adult targets from one another. A few exceptions exist; for instance, for both the MCI and Procrustes measures, there is no overlap in notches between the liquids /ɹ/ and /l/ and the next most complex phoneme, suggesting that these sounds are produced with significantly greater tongue shape complexity than other phonemes. Across all three metrics, vowels were generally associated with low values, and /ɹ/ and /l/

were generally associated with high values. *Figure 2.2* also shows that rank ordering mostly agreed with the Dawson categories, in that most of the phonemes with the lowest medians belong to the low-complexity category and most of the phonemes with the highest rank belong to the high-complexity category. A notable exception is that /ɑ/ and /ʌ/ were considered low and /w/ and /j/ were considered medium complexity according to the Dawson categories, but /ɑ/ and /ʌ/ were ranked relatively high by NINFL /w/ was ranked relatively low by Procrustes and relatively high by MCI, and /j/ was ranked relatively high by NINFL. Additionally, /ʒ/ and /d/ were considered high complexity based on the

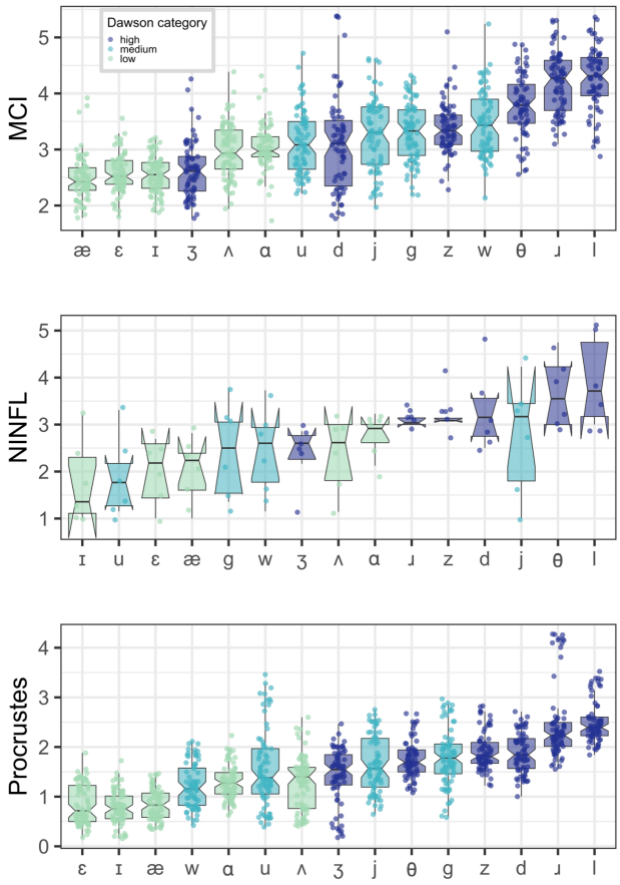

Figure 2.2. Adult targets by MCI, NINFL, and Procrustes, colored by Dawson categories.

Dawson categories, but MCI and NINFL placed /ʒ/ among the lower ranked phonemes while MCI ranked /d/ in the middle of the range.

For our second objective looking at tongue shape complexity patterns across phonemes in the child data set, *Figure 2.3* shows the same rank ordered notched boxplots as used for adults, where the median-based rank ordering is shown for MCI and Procrustes. For NINFL, since there were multiple word contexts for each target phoneme, the mean value for each speaker by word was calculated in order to break ties between distributions with the same median. As with adults, the plots are rank ordered by the median of these means and notches represent the confidence interval around those medians. As before, horizontal jitter is added to all three plots while vertical jitter is added to the ordinal NINFL plot to avoid overlapping values. To a greater degree than observed in the adult data, the notches overlap across phonemes and do not separate child targets from one another. Based on NINFL only, there is minimal overlap in notches between the liquids /ɹ/ and /l/ and the next most complex phoneme, suggesting that these sounds are produced with significantly greater tongue shape complexity than other phonemes. Among the discrepancies within the child data, /w/ was ranked as higher complexity than expected based on all three metrics. Additionally, even though /l/ and /t/ belong to the Dawson high complexity category, they were the first and third lowest-ranked phonemes based on MCI, while /t/ was the second lowest-ranked phoneme based on Procrustes, and a mid-ranked phoneme based on NINFL. As expected, /l/ was ranked high based on Procrustes and NINFL, while /ɹ/ was ranked among the highest according to all three metrics.

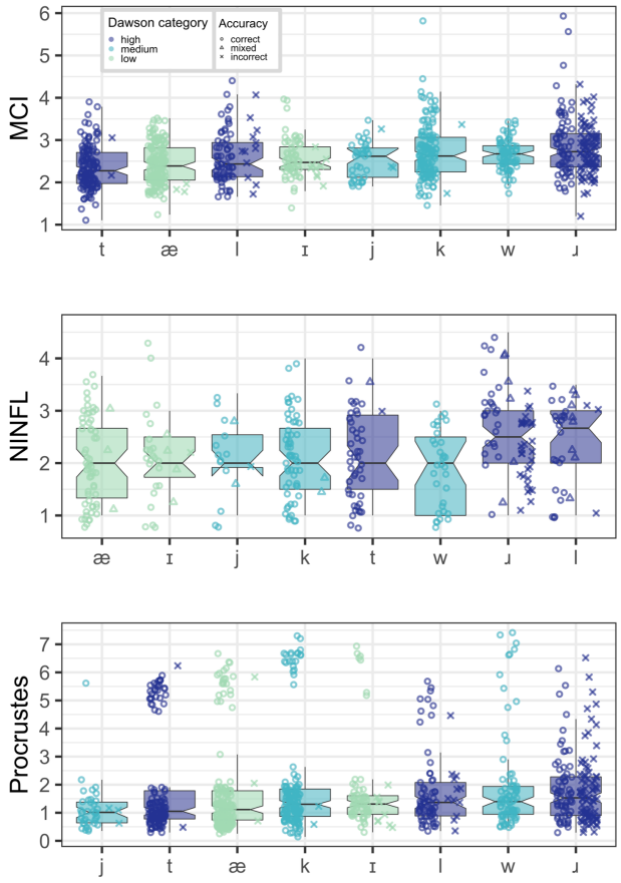

Figure 2.3. Child targets by MCI, NINFL, and Procrustes, colored by Dawson categories.

Because transcription-based accuracy ratings were available for all child productions, the plots show whether each production was considered “correct” or “incorrect”. For the plotted mean NINFL values for each subject and target, an intermediate “mixed” accuracy rating category is included when all productions for that subject/target did not agree. As can be seen in the plot, the children produced most vowels, glides, and stop consonants with a high degree of accuracy. Recall that /l/ (as well as /ɹ/) were produced with less than 90% accuracy across the children in this data set. For /l/, the incorrect productions do not appear to separate from the correct productions based on any metric,

suggesting that accuracy does not mediate the unexpected ordering for /l/ based on MCI. However, for /ɹ/, the incorrect productions do appear lower than the correct productions based on both MCI and NINFL. This suggests that accurate /ɹ/ productions would show even greater separation from the other phonemes than presently shown.

Discussion

The present study explored the utility of ultrasound-derived measures of tongue shape complexity to characterize speech sounds produced by a group of adult speakers and a group of child speakers. Our overall objective was to determine which metrics, taken individually, best represented degree of tongue shape complexity in children as well as adults. Our first objective was to determine whether phonemes or natural classes of phonemes patterned differently with respect to measures of tongue shape complexity for adults. Overall, MCI, Procrustes, and NINFL yielded values that broadly agreed with the Dawson complexity categories. At the individual phoneme level, values of complexity measures were typically low for vowels and high for /ɹ/ and /l/. However, there were also unexpected results, such as the discrepancies found between certain metrics and the expected Dawson categories for /ɑ/, /ʌ/, /j/, /w/, /ʒ/, and /d/ in the adult data set. As hypothesized, for all three metrics in the adult sample, glides had higher tongue complexity than vowels, and liquids had higher tongue complexity than glides. Our second objective was to determine whether the patterns found with adults could also be observed in a sample of TD children for whom some phonemes were still emerging. We found that there was substantial

overlap of tongue shape complexity values across phonemes for MCI, Procrustes, and NINFL. Tongue shape complexity values did not show an exact correspondence with the Dawson categories, primarily due to the high MCI and Procrustes values versus the low NINFL values obtained for /w/, the low MCI values obtained for /l/, and the relatively low values obtained for /t/. Recall from *Table 2.1* that even though the alveolar stop was categorized as high complexity according to the Dawson categories, it was considered low complexity according to the other four taxonomies. For MCI only, the hypothesized relationship was observed between phoneme classes where glides had higher tongue shape complexity than vowels and liquids had higher tongue shape complexity than glides. For NINFL, liquids had higher tongue shape complexity than vowels and glides, but there was no difference observed between vowels and glides. When considering transcription-based accuracy, /ɹ/ showed lower values for some incorrect productions relative to the correct productions based on both MCI and NINFL. Overall, these results suggest that tongue shape complexity measures (MCI, and to a lesser extent, NINFL) may be instrumental in revealing differences between phonemes and phoneme classes for both adults and children, but that these measures do not always accord perfectly with phonetically informed expectations.

The current analyses serve as a proof of concept for how objective tongue shape complexity measurements can be applied to child data. Despite differences in size between adults and children, it can be observed that the ranges of values for both MCI and NINFL are similar, highlighting how these metrics function

independently from vocal tract size. However, the Procrustes measure was associated with different ranges of values for the adult sample (0-5 units) and child sample (0-8 units), suggesting that this measure may be less optimal for comparisons across different-sized vocal tracts. It can also be observed that earlier-developing phonemes such as /æ/ and /ɪ/ have MCI, Procrustes, and NINFL values that are roughly the same for adults and children, whereas later developing phonemes such as /l/ and /ɹ/ appear to have notably higher values for MCI and NINFL (and to a lesser extent, Procrustes) in adults than in children. Although it is not possible to make direct quantitative comparisons between adults and children in the present study due to the different tasks used to collect the target sounds in the two samples, the qualitative differences between adult and child tongue contours for these sounds support the hypothesis that tongue shape complexity increases with age for certain phonemes, and that these differences may be detectable over the course of maturation. It remains unknown whether children's tongue contours continue to show reduced complexity after production becomes perceptually accurate or if reduced tongue shape complexity would persist covertly. Future research should determine at what point in childhood tongue shape complexity becomes adultlike, and how this trajectory relates to changes in perceived accuracy.

Above we noted that the Procrustes range for the child productions was larger (0-8) than the range for the adult productions (0-5), suggesting that this measure may be less optimal for analyses of speakers with varying vocal tract lengths. It is also important to consider methodological differences between the

Procrustes measure as compared with MCI and NINFL. MCI and NINFL both rely on the degree of curvature at each point along the contour (the combined curvature of all points for MCI and the number of points with curvature above a set threshold for NINFL) to provide a quantitative representation of complexity. The Procrustes metric is understood to rely heavily on what resting shape was selected as the starting point for the subsequent translation, rotation, and scaling. However, there is no agreed-upon method for eliciting a resting contour, and the resting contour could in principle differ in complexity between individuals of different sizes. That is, for younger speakers, the tongue fills the oral cavity more completely, so a resting shape for a younger speaker might track the palate more closely than in larger/older speakers with more space in the oral cavity. Due to this possibility, it follows that Procrustes-based complexity values may not be comparable across individuals. Synthesizing across these considerations, we favor MCI and NINFL because they are the two metrics that quantified degree of curvature most directly and with equivalence across vocal tracts of different sizes.

We now reflect on the relative performance of these two preferred measures, MCI and NINFL, in dividing adult and child data by phoneme classes. As seen in *Figure 2.2* for adults and in *Figure 2.3* for children, separation across phonemes was relatively limited for both metrics. Especially for children, there was substantial overlap of the notched boxplot intervals across targets, with mean/median values near 2 for virtually all phonemes in both cases. Considering the ranking ordering of phonemes based on mean/median complexity, a handful of sounds exhibited the predicted behavior in the adult data set: /æ/ and /ɪ/

contours were consistently among the lowest in complexity; /g,k/ contours were consistently intermediate; and /ɹ/ and /l/ contours were consistently high in complexity. However, /j/, /l/, /w/, /ʒ/, and /d/ were not classified consistently across metrics in both data sets. For adults, /j/ was identified as high complexity based on NINFL but mid-complexity based on the other two metrics, while for children, /j/ was ranked as the lowest complexity phoneme based on Procrustes. For children, /l/ was identified as high-complexity by NINFL and Procrustes, but was among the lowest-complexity targets based on MCI. For adults, based on NINFL, /d/ was ranked relatively higher complexity compared to NINFL, and based on both MCI and NINFL, /ʒ/ was ranked lower in complexity than expected, while /w/ was ranked higher in complexity than expected.

Finally, we reflect on methodological differences between MCI and NINFL as they relate to the discrepancies between these measures observed in both the adult and the child samples. Notably, /l/ was characterized by high complexity based on NINFL but low complexity based on MCI in the child data set only. Additionally, /d/ was characterized by relatively high complexity in the adult sample based on NINFL but not MCI, while /t/ was characterized by relatively low complexity in the child sample based on both metrics. The relative complexity of /w/ was reversed across the two measures: MCI indicated relatively high complexity for /w/, whereas NINFL categorized /w/ with relatively low complexity. This finding was especially pronounced in the adult data. Recall that MCI is driven by curvature and NINFL is driven by the number of inflections; as such, MCI is higher when the size of the local curvature is low (as with the locally

tight curvature of the tongue tip that occurs in a retroflex /ɹ/), whereas NINFL is not sensitive to differences in curvature size. These computational differences may account for some of the discrepancies observed across metrics. See Kabakoff et al. (2021) and Kabakoff et al. (submitted) for more discussion of the differences between these metrics.

When considering some of the unexpected findings in the present study, it is important to note that the metrics we describe are limited to a midsagittal section of the tongue. In many cases, looking at multiple sections of a tongue shape would reveal complexity that cannot be reflected in a single midline section of the same tongue shape. This may be especially relevant for rhotics and sibilants (such as /ʃ/ and /ʒ/), which are produced with a midline groove with bracing of the lateral edges of the tongue. This suggests that a combination of sagittal and coronal sections, or three-dimensional ultrasound imaging, may be necessary for a comprehensive characterization of the tongue shape complexity associated with different phoneme classes. This may be particularly true in adults for the postalveolar fricative that was ranked lower in tongue shape complexity than expected based on both metrics. That is, access to a coronal section may have revealed parasagittal complexity for such targets with lateral bracing. Likewise, as presented in Gibbon (1999) and discussed in Kabakoff et al. (2021), EPG has revealed that mature /t/ is produced with both an alveolar constriction and lateral bracing, suggesting that the complexity for this target sound also cannot be fully represented in the midsagittal section. In addition to exploring other anatomical sections of the tongue, alternative measures of ultrasound tongue shapes or new

articulatory technologies may be needed to better quantify tongue shape complexity in a manner that will reliably differentiate productions based on phoneme, age or disorder status of the speaker, and accuracy rating. Discrepancies between expected and observed complexity may also reflect noise generated from the specific frame that was analyzed. That is, tongue contours in frames just before versus at the point of maximal constriction might differ in tongue shape complexity but not reflect meaningful differences in motor skill. Using a higher frame rate is recommended in order to increase the likelihood of capturing the actual point of maximal constriction.

The present study provides a foundation for using ultrasound-based metrics of tongue shape complexity to characterize speech productions in children as well as in adults. Although any single token of a speech sound is unlikely to be accurately classified based on any of the present tongue complexity metrics, the current analyses provide a strong case that vowels have lower tongue shape complexity than liquids. This work represents an extension of the methods used in Dawson et al. (2016) to a child population, as well as an extension of what was previously observed with EPG to the relatively more affordable and accessible ultrasound technology. Establishing sensitive and valid metrics of tongue shape complexity could make a substantive contribution to a future understanding of how motor factors influence the course of speech development in children. Although the current study did not quantitatively control for degree of perceived accuracy of the children's productions, the wide notched intervals for the late-developing sounds in our child data suggest that both differentiated and

undifferentiated gestures may have been represented in TD children's productions of these targets. In addition to Kabakoff et al. (2021) and Kabakoff et al. (submitted), further research should determine whether tongue shape complexity for these targets distinguishes perceptually correct from incorrect productions, as found for rhotic targets in Preston et al. (2019).

As an additional future direction, it would be valuable to examine whether lingual differentiation is higher for TD children than children with SSD, as found in Gibbon (1999), Preston et al. (2019), and Kabakoff et al. (2021). Subsequent research could investigate whether tongue shape complexity measures can identify subtypes within the population of children with SSD, such as those who are most likely to show errors that persist later in development or those whose speech errors are most likely to have a motor-based etiology. This could pave the way for a clinical application in diagnosis and treatment planning. That is, if tongue shape complexity measures from a child with SSD are suggestive of motor involvement (i.e., reduced tongue shape complexity for late-developing phonemes), a motor-based approach to treatment, such as ultrasound biofeedback (e.g., Bernhardt, Gick, Bacsfalvi, & Adler-Bock, 2005; Cleland, Scobbie, & Wrench, 2014), might be recommended. If tongue shape complexity measures do not provide evidence for motor involvement (i.e., relatively high tongue shape complexity), a phonological approach to intervention might instead be recommended. For those with motor-based impairments, the current metrics may also serve the additional purpose of helping to quantify a baseline level of tongue shape complexity and track progress over the course of treatment.

Conclusion

In the present study, we asked which among three metrics best reflect the degree of complexity of a given tongue shape. Results from applying MCI, NINFL, and Procrustes metrics to adult productions suggested that they group the contours broadly into the complexity categories proposed by Dawson et al. (2016), although there were exceptions. These metrics also can be applied to child productions, potentially providing insight into developmental patterns that are not observable through perceptual analyses alone. Our evidence suggests that MCI and (to a lesser extent) NINFL are well-suited for detecting differences in tongue shapes, whereas the Procrustes method poses additional analytical challenges. In order to determine the true utility of these metrics in clinical populations, future research should apply these metrics to child populations differing in age, disorder status, and degree of perceived accuracy.

Acknowledgments

We thank Katherine M. Dawson for sharing the data that formed the basis of the adult analyses, Graham Tomkins Feeny for coordination of measurement in GetContours, and Emily Phillips, Megan Leece, and Twylah Campbell for data collection at the three sites. We acknowledge Siemens Medical Solutions USA, Inc., for making their Acuson ultrasound scanner available. This research was supported by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under Grant F31DC018197 (H. Kabakoff, PI), Grant R01DC013668 (D.H. Whalen, PI), and Grant R01DC017476 (T. McAllister, PI). Stimuli from Dawson et al. (2016) were created with support

under Grant R01DC002717 (D.H. Whalen, PI). Additional support was provided through an Acoustical Society of American Stetson Scholarship and an American Speech-Language-Hearing Foundation New Century Scholars Doctoral Scholarship.

MANUSCRIPT 2: Extending ultrasound tongue shape complexity measures to speech development and disorders

Abstract

Purpose: Generalizations can be made about the order in which speech sounds are added to a child's phonemic inventory and the ways that child speech deviates from adult targets in a given language. Developmental and disordered speech patterns are presumed to reflect differences in both phonological knowledge and skilled motor control, but the relative contribution of motor control remains unknown. The ability to differentially control anterior versus posterior regions of the tongue increases with age, and thus complexity of tongue shapes is believed to reflect an individual's capacity for skilled motor control of speech structures.

Method: The current study explored the relationship between tongue complexity and phonemic development in children (ages 4-6) with and without speech sound disorder producing various phonemes. Using established metrics of tongue complexity derived from ultrasound images, we tested whether tongue complexity incrementally increased with age in typical development, whether tongue complexity differed between children with and without speech sound disorder, and whether tongue complexity differed based on perceptually rated accuracy (correct versus incorrect) for late-developing phonemes in both diagnostic groups.

Results: Contrary to hypothesis, age was not significantly associated with tongue complexity in our typical child sample, with the exception of one association between age and complexity of /t/ for one measure. Phoneme was a significant predictor of tongue complexity, and typically developing children had more complex tongue shapes for /ɹ/ than children with speech sound disorder. Those /ɹ/

tokens that were rated as perceptually correct had higher tongue complexity than the incorrect tokens, independent of diagnostic classification.

Conclusion: Quantification of tongue complexity can provide a window into articulatory patterns characterizing children's speech development, including differences that are perceptually covert. With the increasing availability of ultrasound imaging, these measures could help identify individuals with a prominent motor component to their speech sound disorder, and could help match those individuals with a corresponding motor-based treatment approach

Introduction

During the process of speech maturation, a child must acquire perceptual-phonological knowledge and articulatory-motor control. In the phonological domain, the child must learn which sounds are contrastive (in perception and production) in a given language while acquiring language-specific knowledge about positional constraints and allophonic alternations affecting those phonemes (Boersma & Hayes, 2001; Tesar & Smolensky, 1998). At the same time, a child must also identify, refine, and organize the articulatory gestures used to achieve the desired phonetic output (e.g., Namasivayam et al., 2020; Noiray, Abakarova, Rubertus, Krüger, & Tiede, 2018; Noiray, Ménard, & Iskarous, 2013; Studdert-Kennedy & Goldstein, 2003). Thus, both phonological knowledge and skilled control of speech structures are required to achieve perceptually accurate production. Comparing a child's speech development to normative values can reveal delays or abnormal patterns associated with speech sound disorder (SSD), which reflects a combination of the state of a child's phonological knowledge and

the maturation of motor control of their speech structures. However, the relative contributions of perceptual, phonological, and articulatory-motor abilities are not fully understood, and it is likely that they interact and cannot be fully separated (Cleland et al., 2017; Fey, 1985; Gibbon, 1999; McAllister Byun & Tessier, 2016). The purpose of the current study is to develop measures that can highlight the relative contribution of motor maturation to the development of adultlike speech patterns in children with and without SSD. If the relative contribution of motor development can be quantified as distinct from phonological or perceptual development, then children with primarily phonological versus motor-based impairments could be identified and matched with a corresponding treatment approach.

Lingual differentiation

In early stages of development (Fletcher, 1989), as well as in disordered speech (Gibbon, 1999), a reduced ability to isolate control of anterior versus posterior regions of the tongue is likely to play a role in children's nonadultlike speech patterns. This capacity for different regions of the tongue to operate semi-independently is referred to as "lingual differentiation"; gestures that lack a typical degree of independent lingual control may be described as "undifferentiated gestures" (Gibbon, 1999). Fletcher (1989) compared patterns of linguopalatal contact between children ages 6-14 using static palatography, a method in which articulatory contact is tracked by putting ink on the tongue and observing its transfer to the palate after a targeted utterance. For alveolar stop targets, Fletcher (1989) found that there was more linguopalatal contact for

younger children than older children, and the amount of contact decreased incrementally across age groups (6-8 years, 9-10 years, 11-14 years). More specifically, this pattern of fewer palatal sensors activated was characterized by more posterior contact for velar targets and wider grooves for sibilant consonants. This pattern suggests that lingual differentiation is a motoric skill that increases in a gradual fashion over the course of maturation. While the relative size of the tongue to the vocal tract does continue to decrease into early childhood at the same time that differential lingual control develops (e.g., Bosma, 1963), these anatomical changes are likely to work synergistically with the refinement of somatosensory feedback and lingual motor control (Gibbon, 1999).

Based on numerous case studies, it has been argued that undifferentiated lingual contact occurs more often in populations with atypical speech than in children with typical speech (Gibbon, 1999; A. Lee, Gibbon, & O'Donovan, 2013), though well-controlled studies are needed to support this assertion. Gibbon (1999) reviewed nine studies that reported data collected using electropalatography (EPG), a method akin to palatography but using electrodes encased in a pseudo-palate. From a sample of 17 children ages 4-12 with SSD, Gibbon (1999) observed that 71% showed undifferentiated gestures that involved broad linguopalatal contact exceeding what is typical in adult production. Specifically, most of these children were observed to produce alveolar stop targets with lingual contact that spanned both alveolar and palatal/velar regions, as opposed to localized contact on the alveolar ridge. Within children exhibiting undifferentiated gestures, she also drew a distinction between those with

"discrete" lack of differentiation (affecting only a specific speech target), and "widespread" lack of differentiation (affecting many speech targets). For children with undifferentiated gestures, and especially for children with widespread undifferentiated gestures, reduced lingual differentiation is thought to represent a motor simplification strategy associated with reduced control over the different functional regions of the tongue. Typically developing (TD) children try out various gestures and configurations and through repeated production, eventually develop the ability to isolate control of different lingual regions in order to achieve the desired phonetic output (Guenther, 2016; Schwartz, Basirat, Ménard, & Sato, 2012). Likewise, children with SSD may also be able to overcome motor-based constraints through a combination of maturation and targeted therapy.

It is important to mention that the degree of lingual differentiation for an individual production cannot necessarily be predicted on the basis of transcription or perceptually rated accuracy (Gibbon, 1999). Demonstrating this point with EPG, Gibbon et al. (1993) reported data from a nine-year-old girl with SSD who exhibited a pattern of backing in which alveolar stop consonants were transcribed with a velar place of articulation. However, EPG revealed that the participant's alveolar stops were realized with a pattern of simultaneous velar and palatal or alveolar contact that was distinct from her gestures for velar targets, making this an instance of "covert contrast" (Gibbon et al., 1993). McAllister Byun et al. (2016) used ultrasound to detect covert contrast in a preschool-aged child's pattern of velar fronting, where velar targets were perceptually neutralized with alveolar targets. This pattern was characterized by substantially higher dorsum

excursion in velar targets than in alveolar targets, even though the two were perceptually neutralized. Ultrasound has also been used to reveal "covert errors," or atypical lingual configurations that are not detected perceptually (Cleland et al., 2017; Klein et al., 2013). For example, Cleland et al. (2017) observed covert errors in which a school-aged child with SSD produced perceptually accurate /k/ using abnormal retroflexion. Taken together, these cases of covert patterns of lingual differentiation suggest that close inspection of such articulatory patterns may offer information about a child's state of speech development, particularly in the area of motor maturation, that cannot be obtained from transcription or perceptual ratings alone.

Why lingual differentiation matters

Lingual differentiation matters not only because it could serve as a window into motor development, but because it may be key to determining which treatments to recommend for children with SSD. Clinicians treating these children are confronted with the option of choosing among different treatment approaches, notably a traditional articulatory approach versus a phonological approach to treatment. Traditional motor-based treatment approaches focus on repetitive production practice, building from producing the target sound in isolation to producing it at the conversational level (Van Riper, 1978). Such approaches often incorporate principles of motor learning (e.g., manipulating the frequency and type of feedback provided) in order to facilitate maintenance and generalization of targeted speech skills (Maas et al., 2008). In contrast, phonological treatment approaches consider how children's errors can be categorized according to

phonological processes or distinctive features, then aim to reorganize the child's phonological system through the elimination of inappropriate phonological patterns (Hodson & Paden, 1983), the introduction of absent distinctive features (McReynolds & Bennett, 1972), or the realization of phonological distinctions in minimal pairs (Barlow & Gierut, 2002). Phonological approaches also focus on building the child's ability to perceive phonetic differences and relate them to meaningful contrasts in their language (Rvachew & Brosseau-Lapré, 2012). Finally, Rvachew and Brosseau-Lapré (2015) highlight the importance of building phonological awareness and vocabulary when treating children with phonological disorders. In principle, different children should be paired with different intervention approaches based on their relative strengths and weaknesses in perceptual, motoric, and phonological domains. In practice, however, speech-language pathologists often find it challenging to determine which approach is most suitable for a given child.

Gibbon (1999) and Cleland et al. (2017) have argued that information available through articulatory imaging could be of value when selecting a treatment approach for a given child. That is, children who exhibit speech errors but show typical articulatory patterns (i.e., normal degree of gestural differentiation, absence of covert errors) may benefit from treatments that emphasize phonological approaches. In contrast, for children who do exhibit atypical undifferentiated gestures or covert errors, treatments that foreground the motor-based component of speech might be most beneficial. Furthermore, among those children with undifferentiated gestures, Gibbon (1999) predicted that those

with discrete undifferentiated gestures would respond more rapidly to motor-based approaches than those with more severe articulatory-motor delay, characterized by widespread undifferentiated gestures. More research is needed to understand the relationships between an individual's degree of lingual differentiation and likely response to different treatment approaches, but this is an area with strong potential to inform clinical decision-making.

Lingual differentiation across phonemes

Some phonemes inherently involve a greater degree of lingual differentiation than others; phonemes that require a greater degree of lingual differentiation are thought to be acquired later than those that are produced with simpler shapes. Cross-linguistic reports reveal a broad pattern in which plosives, nasals, nonpulmonic consonants, and speech targets with posterior place of articulation are typically acquired before fricatives, affricates, trills, flaps, and speech targets with anterior place of articulation (McLeod & Crowe, 2018). A recent systematic review of fifteen studies including over 18,000 children acquiring American English reported that plosives, nasals, and glides were mastered by age four, affricates were mastered by age five, liquids were mastered by age six, and fricatives were mastered by age seven (Crowe & McLeod, 2020). Also looking at English-speaking children only, Studdert-Kennedy and Goldstein (2003) reported a taxonomy of phonetic development in which young children first showed mastery of voiceless stops, nasals, glides, and /h/, and then later added voicing contrasts as they developed the ability to coordinate laryngeal gestures with lingual gestures. A third stage occurred when children were able to

coordinate jaw height and/or degree of constriction with lingual gestures in order to produce fricatives and affricates. They described the final stage as occurring when multiple lingual constrictions began to occur, allowing for the production of such liquid consonants as /l/ and /ɹ/. Overall, while factors such as perceptual salience are also relevant, previous research has established that the capacity for discrete control of different lingual regions is an important factor in shaping developmental schedules of consonant acquisition.

For late-developing, complex sounds like laterals and rhotics, motor simplification strategies have been observed in productions by both TD and SSD populations. Using ultrasound imaging, Lin and Demuth (2015) observed that young children tended to produce laterals with a single constriction, with the rate of simplification decreasing with age from 3-7. Interestingly, they noted that onset /l/ with a single constriction was generally perceived as an accurate light [l], while gesturally simplified productions of coda [ɬ] were not accepted as accurate. Since both onset and coda /l/ are produced with two constrictions in adult English, the young children's simplified [l] in onset position might be considered a case of covert error. For rhotic targets, TD children's error patterns reflect various strategies to simplify the required complex contour, including omissions of either the anterior or posterior constriction, or combining both constrictions into one centralized undifferentiated constriction (Gick et al., 2007). The resulting articulation may be perceived either as a substitution of another phoneme (e.g., [w]) or as a subphonemic distortion. Klein et al. (2013) used ultrasound to explore the relationship between lingual contours, acoustics, and perceptual accuracy in

rhotics produced by three TD children (ages 5-11) and two children with SSD (ages 5 and 6). For children with SSD, they found that greater perceptual accuracy was associated with more differentiated gestures, and this relationship was also reflected in acoustic measures. Finally, for children with SSD, degree of lingual differentiation was observed to increase over the course of treatment targeting rhotics.

Measuring tongue complexity

The preceding discussion suggests that measuring degree of tongue shape complexity could be valuable for understanding developmental speech patterns, as well as for differential diagnosis and treatment planning in the context of SSD. It is thus important to consider different approaches that can be used to extract measures of tongue shape complexity. EPG is useful because of its ability to reveal whether linguopalatal contact is localized to discrete regions or spans multiple areas. However, use of EPG requires an individually customized palatal prosthesis. Although there is evidence of efficacy supporting the use of EPG for both diagnosis and treatment (Bernhardt et al., 2003), its clinical utility is limited by the high cost and time delay required to fit and manufacture the customized device. Additionally, the approach is not considered well-suited for young children with growing vocal tracts, who may require several palates over the course of intervention (Gibbon, 1999). Therefore, there is a need for alternative approaches to measuring tongue complexity that are better suited for young children.

In contrast to EPG, ultrasound imaging is increasingly accessible and affordable, minimally invasive, and ready to use, even with small children. Although it is not equipped to directly show palatal contact, ultrasound does provide a continuous midsagittal lingual contour, revealing potentially relevant information about tongue shape. Like EPG, ultrasound has also been used to show fine detail about child speech patterns that may not be detected perceptually, including covert errors and covert contrasts (Cleland et al., 2017; McAllister Byun et al., 2016). To understand how EPG measures of lingual differentiation might relate to ultrasound measures of tongue complexity, consider the contrast between a [k] sound produced with undifferentiated linguopalatal contact (i.e., spanning from alveolar to velar regions of the palate) versus a [k] sound produced with linguopalatal contact isolated to the velar region. In the former case, the contour of the tongue will track the palate and therefore is likely to have a simple curved shape. By contrast, production of differentiated velar contact requires actively lowering the front of the tongue, which will generally produce a more complex shape in which the anterior and posterior tongue are separated by a change in curvature. Likewise, Gibbon (1999) reported that a mature /t/ is produced with an alveolar constriction and lateral bracing that appear as a "spoon-like" contour through EPG. The raised and lowered portions of such a configuration would also contribute to increased complexity of the tongue contour as seen through ultrasound, although some of the complexity would be apparent only in coronal rather than midsagittal section. Despite this reasoning that EPG and ultrasound measures are likely to be related to one another, we acknowledge

that simultaneous ultrasound and EPG (or probe-stabilized ultrasound and hard palate traces) would be required in order to determine whether the undifferentiated lingual gestures reported by Gibbon (1999) indeed have lower tongue shape complexity. Of course, there are also differences between EPG measures of differentiation and ultrasound measures of tongue complexity. Most notably, ultrasound can be used to capture the shape of the tongue during speech sounds produced with limited palatal contact, such as the liquids /l/ and /ɹ/. Given these similarities, the present study explored the possibility that ultrasound may be valuable as a means to quantify tongue complexity, which may prove to be a relatively more accessible source of insight into a child's stage of motor development than EPG.

Numerous measures have been proposed to quantify the complexity of tongue contours extracted from ultrasound images (Dawson et al., 2016). Kabakoff, Beames, Tiede, Whalen, and McAllister (under review) compared three analytical approaches for the quantification of the degree of complexity of tongue shapes from ultrasound images and found that the Modified Curvature Index (MCI; Dawson, Tiede, & Whalen, 2016) and the number of inflection (NINFL) points (Preston, McCabe, Tiede, & Whalen, 2019) are the metrics that best correspond with pre-established tongue complexity classes. MCI for a given tongue contour is determined by first computing the absolute curvature (the reciprocal of the tangent circle) at each point of a normalized equidistant set of points describing it and then integrating. MCI was found to be higher in phonemes with a tongue tip constriction or with multiple constrictions (including

/ɹ/ and /l/) than in sounds with a single constriction of the tongue body (including vowels such as /æ/ and /ɪ/) in a sample of adult speakers (Dawson et al., 2016). NINFL represents the number of sign changes in filtered curvature along the arc of a given tongue contour exceeding a threshold, where each change from concave to convex or vice versa is considered an inflection (Preston et al., 2019). NINFL values have been found to be higher in /ɹ/ contours from TD children than in children with SSD, higher in perceptually accurate /ɹ/ tokens than in perceptually inaccurate /ɹ/ tokens, and higher after treatment for /ɹ/ than before treatment (Preston et al., 2019). This evidence suggests that both MCI and NINFL may be useful in distinguishing tongue contours of children based on various factors such as age, disorder status, and/or perceptual accuracy. Furthermore, because these measures are derived from quantifying curvature, they are invariant with respect to translation, scaling, and rotation factors and are thus appropriate for comparison within and across children with different sized vocal tracts and under different imaging conditions.

The current study

The present study investigated whether there are differences in degree of lingual differentiation, as measured with ultrasound, across developmental and diagnostic groups. To do this, we used the metrics of complexity selected previously (MCI and NINFL) to ask three questions about the extent to which tongue complexity varies within and across children of different clinical presentations. The first question asked whether tongue complexity in TD children increases with age, and whether there are any phoneme-specific patterns in this

relationship. We did not expect to find developmental changes in tongue complexity among the earlier-developing phonemes, including the reference vowel and glides, but we did predict that tongue complexity would increase with age for later-developing phonemes such as /ɹ/ and /l/. The second question was whether TD children aged 4-6, who may or may not exhibit developmental speech errors, use more complex contours than peers with SSD. Identification of such a difference between TD and SSD groups would suggest that ultrasound, like EPG, can be used to identify differences in lingual differentiation between TD and SSD groups. Our third question was whether perceptually incorrect /l/ and /ɹ/ productions are associated with less complex tongue contours than perceptually correct productions of the same targets. This pattern was found by Preston et al. (2019) for school-aged children with SSD, so we sought to determine whether this pattern holds for younger children with SSD, as well as for same-age TD children who exhibit some developmentally appropriate incorrect productions. If ultrasound tongue contours are sensitive to differences in tongue complexity within and between the young children in the present sample, these measures may be diagnostically useful as a way to classify young children's errors as more phonological or motoric in character, which may ultimately facilitate pairing children with optimal intervention approaches, as suggested in Gibbon (1999) and Cleland et al. (2017). However, it is also important to acknowledge at the outset that our participants were younger than the school-aged children with residual speech errors who formed the basis of the studies that informed our research

questions (Cleland et al., 2017; Gibbon, 1999), and the nature of diagnosis and treatment planning may look different across these distinct groups.

Methods

Participants

The data set included measurements from 24 children who participated in an evaluation at one of three sites, including Haskins Laboratories, Molloy College, and Syracuse University. An additional participant was excluded because the frame rate for the ultrasound video was 12 frames per second, which was substantially lower than the other files. The participating children had a mean age of 5;1 (years; months, 4;0-6;3) and included 12 girls and 12 boys. Seven children were diagnosed with SSD based on the evaluation procedures described below. The other 17 children had no history of speech or language impairment and passed a pure-tone hearing screening at 500, 1000, 2000, and 4000 Hz at 20 dB HL. For many of these TD children, /ɹ/ and /l/ were not always produced in an adultlike fashion, as these phonemes were emerging along a developmentally appropriate trajectory. The small difference in age between the two groups was non-significant (mean of SSD group = 58.00 months; mean of TD group 62.59 months, $t(10.49) = -1.16$, $p = 0.27$). See *Table 3.1* for details about the children, including site of evaluation, age, gender, and diagnostic classification.

The children participated in a two-day evaluation beginning with a hearing screening, an oral mechanism examination, and an introduction to the ultrasound machine. Speech and language measures for all participants included a conversational play language sample, the Hodson Assessment of Phonological

Patterns-3rd Edition (HAPP-3, Hodson, 2004), the receptive language tasks in the Clinical Evaluation of Language Fundamentals Preschool-2 (CELF-P2, Wiig, Secord, & Semel, 2004), and the Speech Assessment and Interactive Learning System (SAILS) perceptual measure (Rvachew, 1994). For the present analysis, the criterion for inclusion in the SSD group was a standard score at or below 80 on the HAPP-3. The children between the ages of 4;0 and 5;11 who met this criterion and exhibited at least three phonological patterns that were applied in at least 40% of contexts on the HAPP-3 were eligible to participate in a subsequent intervention study involving Cycles treatment. These participants were administered additional tests such as the Syllable Repetition Task (Shriberg & Lohmeier, 2008; Shriberg et al., 2009).[3] Although this test battery was not equipped to identify whether the children classified as having SSD also had speech motor delay, a child's percentage accuracy score from the Syllable Repetition Task in combination with the number of additions transcribed in that task for that child can be used to help with the identification of speech motor planning difficulties (Shriberg, Lohmeier, Strand, & Jakielski, 2012). Complete participant evaluation scores can be found in the supplemental materials.

[3] Based on these criteria, three children (18M, 19F, 20M) were classified in the SSD group and were deemed eligible to receive the subsequent longitudinal course of intervention. Four children (21M, 22M, 23M, 24M) were classified in the SSD group but were not eligible for treatment. All seven children were included in the SSD group for the purpose of the analyses reported here, but those children who were not eligible for treatment were not required to complete the supplemental measures.

Data collection

Ultrasound recordings were collected with a Siemens Acuson X300 with a C8-5 wideband curved array transducer (frequency range 3.1–8.8 MHz, 25.6 mm footprint, 109 degree field of view) at Molloy College; with a Siemens Acuson X300 with a C6-2 wideband curved array transducer (frequency range 1.8–6 MHz, 73.0 mm footprint, 90 degree field of view) at Haskins; and with a Telemed Echoblaster 128 with a PV 6.5 wideband curved array transducer (frequency range 5–8 MHz, 156 degree field of view) at Syracuse University. Scanning settings differed for participants at each site.[4] The ultrasound probe was placed in a microphone stand while the clinician supported alignment of the probe with the child's head. Due to the young age of the participants, the probe was not fixed relative to the child's head (e.g., with a helmet). However, blue dots for automated optical tracking were placed on the child's face and on the probe and used to identify and discard frames featuring an excessive degree of displacement

[4] The ultrasound frame rate and depth varied across sites/participants (Molloy: 43-49 frames per second with 60-70 mm depth; Haskins: 36-37 frames per second with 80 mm depth; Syracuse: 21-25 frames per second with 110 mm depth). Ultrasound video was recorded on a PC at 60 frames per second through an AverMedia video capture card at Molloy and Haskins, and at 35 frames per second with Debut (NCH Software) at Syracuse. Although it is not ideal to have divergent setups across sites, we regard the impact of these differences as relatively minor for the following reasons: a) at even the lowest ultrasound frame rate, the selected frame could be at most 48 ms from the true frame of interest, which is sufficient for the present non-dynamic analysis; b) Although reduced zoom depth may result in fewer pixels available, curvature-based indices of ultrasound data (such as MCI and NINFL) can still be measured without knowledge of spatial orientation (Ménard et al., 2012; Stone, 2005), which would include size differences introduced by depth variation.

of the head relative to the probe, as described below. Children were initially familiarized with pictures used for elicitation prior to placement of the ultrasound. If a child had difficulty naming a picture during ultrasound imaging, the evaluating clinician attempted to elicit the target using semantic cueing or delayed imitation; direct imitation was used only as a last resort. The clinician also monitored the ultrasound image during elicitation, and if they were unsure of the quality of the image for a given trial, an additional production was prompted. At least three repetitions each of 16 words, randomly ordered, were elicited in this fashion. If more than three usable repetitions were produced, the additional productions were retained. Initial consonants included /j/ in "yam", /w/ in "wake" and "wing", /k/ in "cape", "cat", "coat", "key", /t/ in "tape", "tea", and "toe", /l/ in "lake" and "lamb", and /ɹ/ in "rake", "rat", "ring" and "rope." Final consonants were not analyzed. The vowel /æ/ (in "cat", "rat", "lamb", and "yam") was also analyzed for comparison with the target consonants because it was regarded as representing a relatively neutral tongue shape; vowels in other words (which included /i/, /ɪ/, and diphthongs) were not analyzed.[5] See *Table 3.1* for the number

[5] The word list was designed expressly to probe onset lingual singleton consonants that are frequently misarticulated, along with the counterparts that are commonly produced as substitutions for those sounds (e.g., /t/ for /k/, /w/ for /ɹ/, /j/ for /l/). A range of vowel contexts, were utilized in order to maximize imageability of the words for ease of elicitation with our sample of young children. For each word, considerations were made to maximize both imageability and the child's familiarity with the word. This process resulted in selection of words with sounds that may not be the earliest emerging (e.g., "cat" was preferred to the potentially earlier-emerging "cap"). Minimal pairs were included when possible (e.g., "ring"/"wing," "rake"/"lake," "cape"/"tape," "rat"/"cat"). Sibilants were not included because tongue complexity for those sounds would be visible only in a coronal section (i.e., in the presence of lateral bracing).

Table 3.1. Child participant information, including breakdown of all usable tokens by target for each participant.

	Subject	Site	Gender	Age	æ	j	k	t	w	l (% correct)	ɹ (% correct)
TD	01F	Haskins	F	5;6	11	3	10	13	7	5 (100%)	10 (50%)
	02F	Haskins	F	5;11	4	1	5	3	1	3 (100%)	2 (100%)
	03M	Haskins	M	5;6	10	3	12	8	4	4 (75%)	12 (0%)
	04F	Haskins	F	5;6	11	3	9	4	4	4 (100%)	6 (0%)
	05M	Haskins	M	4;2	5	1	1	2	1	2 (0%)	5 (0%)
	06M	Haskins	M	4;11	7	2	10	9	6	6 (67%)	10 (0%)
	07F	Haskins	F	5;11	12	3	12	9	7	6 (100%)	13 (100%)
	08F	Haskins	F	5;9	7	2	5	7	4	2 (100%)	5 (40%)
	09F	Molloy	F	4;3	10	3	8	4	5	5 (60%)	11 (91%)
	10F	Molloy	F	6;0	8	1	12	4	4	3 (100%)	9 (11%)
	11F	Molloy	F	4;9	12	4	8	15	6	4 (100%)	11 (0%)
	12M	Molloy	M	6;3	16	2	15	15	11	11 (100%)	9 (78%)
	13F	Molloy	F	4;3	4	2	3	10	2	5 (60%)	7 (100%)
	14F	Molloy	F	4;5	7		2	5	2	3 (67%)	6 (0%)
	15M	Syracuse	M	5;3	7	3	3	6	6	6 (83%)	6 (83%)
	16F	Syracuse	F	5;5	11	3	11	9	6	4 (100%)	11 (100%)
	17M	Syracuse	M	4;6	12	3	14	9	5	4 (75%)	16 (0%)
SSD	18M	Haskins	M	4;2	16	3	14	9	6	6 (100%)	12 (50%)
	19F	Haskins	F	5;0	8	2	13	7	7	5 (0%)	13 (77%)
	20M	Haskins	M	5;3	1	1	4	4	3	4 (25%)	9 (0%)
	21M	Molloy	M	4;0	6	1	3	5	1	0 (0%)	9 (78%)
	22M	Syracuse	M	5;3	9	1	4	7	3	7 (0%)	9 (0%)
	23M	Syracuse	M	4;0	12	1	4	8	4	9 (89%)	11 (73%)
	24M	Syracuse	M	5;11	12	2	13	6	7	8 (63%)	11 (0%)
					218	50	195	178	112	**116 (67%)**	**223 (41%)**

of tokens of each target phoneme analyzed for each participant after the exclusion of productions that were judged unusable based on criteria described in detail below. For /l/ and /ɹ/, the total number of productions is listed first with the percentage of perceptually correct productions in parentheses. The number of usable tokens varied across individuals, reflecting differences in factors such as the ability to remain still during ultrasound imaging. To account for these inherent differences, we included a random intercept for each child in the models reported.

Ultrasound measurement

All acoustic and ultrasound processing was performed by trained university students who had taken courses in acoustic phonetics and/or general linguistics and had received project-specific training. Following all initial data processing by students, a graduate student with specialized training in phonetic

analysis (henceforth, “student specialist”) assured consistency across sound files and ultrasound files, redoing the processing for any items that did not meet the established standards described below.

For all sound files, the trained students viewed waveforms and spectrograms in Praat (Boersma & Weenink, 2019) in order to mark off each target sound in the time-synced TextGrid file and labeled the relevant intervals by target phoneme. Sonorant intervals were identified as the period of reduced intensity relative to the following vowel, which included most of the formant transition. Vowel intervals were marked differently depending on what consonant preceded them. After a stop consonant, the start of the vowel interval was marked when the waveform became periodic, but after a sonorant consonant, the vowel interval was defined to begin after the offset of the formant transitions of the prior sonorant. The end of a vowel interval was marked at the stop closure when followed by a stop, or near the beginning of the formant transition when followed by a sonorant. Consonant burst locations were marked at the point of a visible spike of energy in the waveform. In order to capture the articulatory constriction before and after the burst, a Praat script was used to generate a window around the location of the burst (55 milliseconds total) in the acoustic record.

Each acoustic interval marked in a Praat TextGrid file (Boersma & Weenink, 2019) was viewed in the time-synced ultrasound video in Matlab (MathWorks Inc., 2000) using GetContours (Tiede, 2020), a program for ultrasound annotation that supports navigation to the first frame within each marked interval. The trained university students were instructed to step within the

marked interval to find the frame toward the center of the interval that most clearly represented the maximum lingual constriction for each production. It was necessary to select frames manually using this process, instead of automatically selecting the frame closest to the midpoint, because many frames were unusable due to the frequent movements that the young children in our sample tended to make while speaking during ultrasound data collection. The trained students were familiarized with ultrasound images representing typical productions of the target phonemes but were also informed that productions could deviate substantially from the target tongue shape (e.g., in cases of a distortion or substitution). After selecting the target frame, the students placed (henceforth, "tagged") sixteen spline anchor points along the underside of the white line visible on the ultrasound image, which were used to generate 100 contour points evenly spaced along the visible contour. A sample set of sixteen tagged anchor points can be seen for correctly articulated /k/, /j/, /l/, and /ɹ/ targets in *Figure 3.1*. The red anchor points define the yellow fitted tongue contour spline in GetContours.

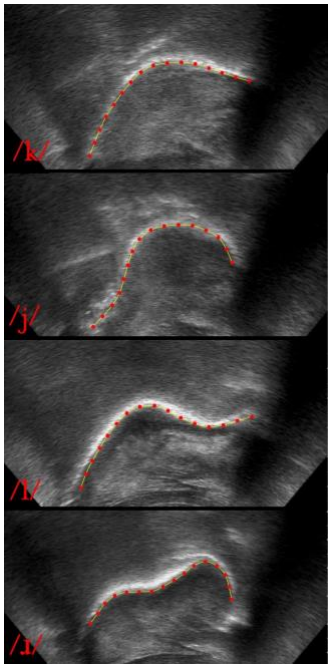

Figure 3.1. Sample sets of sixteen evenly distributed anchor points for four tongue contours traced in GetContours.

Quantitative assessment of ultrasound probe alignment

As noted above, midsagittal ultrasound probe alignment was tracked using blue dots that were placed with adhesive along the vertical midline of the child's face and the ultrasound probe. See *Figure 3.2* for an example of blue dot placement. Video recordings from a front-facing camera were temporally aligned with concurrent ultrasound recordings using cross-correlation of their mutual audio. The alignment of the dots on the child's face relative to the dots on the probe was quantified using an automated in-house Matlab procedure: A line determined by the centroid of the three dots on the forehead and extending through the nose, lip and chin dots (head line) was compared with a line determined by the two or four dots on the probe (probe line) to calculate the lateral displacement (in millimeters) and the angular displacement (in degrees), as

illustrated in *Figure 3.2.* Lateral displacement (⦟LD) was computed as the distance along the perpendicular from the head reference line to the centroid of the probe dots. Angular displacement ($\overleftrightarrow{AD}$) was computed as the counter-clockwise angle from the head to the probe line. Target video frames with more than one standard deviation of lateral displacement (15.4 mm) or one standard deviation of angular displacement (13.3 degrees) across all files were discarded. Across all 24 children, 15.4% (225/1458) of frames were flagged due to angular misalignment and 15.4% (224/1458) of frames were flagged due to lateral misalignment. This resulted in a total of 351 tokens discarded, as 98 of the tokens were flagged for both angular and lateral displacement; a total of 1107 tokens remained after removal of misaligned frames.

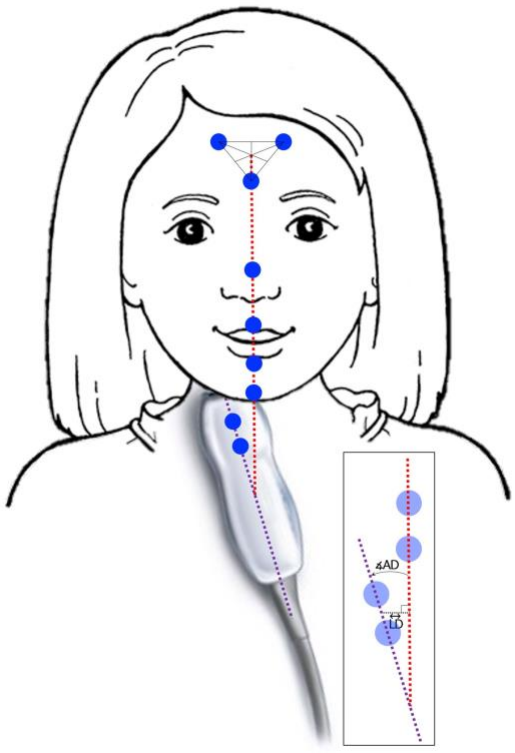

Figure 3.2. Blue dot placement on child's face and on ultrasound probe, illustrating how lateral displacement and angular displacement are defined.

Qualitative assessment of tracked lingual contours

The quantitative analysis described above was intended to identify and exclude the two most common sources of error: lateral displacement of the probe along the right/left axis, and rotation of the probe around the posterior/anterior axis, both of which result in the capture of tongue images which are not midsagittal. Possible rotation around the right/left lateral axis is not tracked by this method, but is irrelevant for the present analyses as the current curvature measures are unaffected by in-plane rotation. However, an additional potential source of error that cannot be tracked using frontal video is probe rotation around the inferior/superior axis; such a misalignment could have the effect of capturing a visual slice not aligned with the midsagittal plane and thus distorting the imaged tongue shape.

In the quality check process, the student specialist confirmed that no selected frames were off-center or unclear. Although an ideal contour would extend from the mandibular shadow to the hyoid shadow, not all contours included both shadows. A frame was considered off-center if the specialist was not confident that anterior and posterior regions of the tongue were adequately visible; an image was considered unclear if the specialist was not confident that the ultrasound was properly tracking the tongue's surface. In such cases, previous and/or adjacent frames were used to inform the tongue's location. In the few cases when a continuous contour could not be visualized for a given frame and the adjacent frames did not assist with revealing the location of the tongue's surface clearly, the token was not tagged and therefore was not included in the analyses.

Therefore, before measurement, all target frames were judged to depict complete and clear contours. Reliability for the above-described frame selection process was quantified for twenty percent of participants (n = 5), and indicated that the distribution of the differences in frame number between the frames selected in the original six files and those selected in the reliability files centered around zero (mean = -0.80 frames, SD = 4.3 frames). For sonorants, the frame difference ranged from 0 to 12 frames, with the exception of one outlier that was drawn from a highly hyperarticulated token in which the /ɹ/ interval was sustained for over 70 frames. For stops, the frame difference ranged from three frames before the original to two frames after the original.

After the specialist deemed all files to be fully and clearly tagged, the 100-point spline contours were exported as x-y image pixel-based coordinates in CSV format for further processing. After extraction, all x- and y-coordinates for all target frames were read into R (R Core Team, 2019) by the first author and scaled (z-score) for both the x- and y-axes within speaker and target in order to adjust for any differences in anterior versus posterior placement of the contour within each image. As a final round of quality assurance of the tagged contours, the first author plotted and visualized all scaled coordinates, and consensus was reached by the first author and student specialist on which contours should be removed as visually anomalous. Thirteen contours were removed because they stood out upon visual inspection as exhibiting high degrees of perseverative coarticulation, as with 10 /æ/ productions that showed multiple constrictions following /ɹ/ (in “rat”). Two tokens of /k/ were excluded (from subjects 21M and 15M) because the

contours did not resemble the shapes of the speakers' other /k/ productions, a phenomenon that was judged to be due to the image not fully capturing the most posterior part of the tongue contour. Additionally, a /w/ contour was removed from 21M because the contour did not resemble the shapes of the speaker's other /w/ productions, which was judged to be due to the image not fully capturing the anterior region of the tongue. After this process of removing outliers based on visual inspection, a total of 1094 tokens remained. Following the criterion introduced in Preston et al. (2019) and followed in Kabakoff et al. (under review), any child token with a NINFL value exceeding five (n = 1) was removed. Similarly, any token with an MCI value exceeding six (n = 1) was removed based on the distribution of values presented in Dawson et al. (2016), leaving a total of 1092 productions in the final data set.

Ratings of perceptual accuracy

All child productions were narrowly transcribed using Phon speech analysis software (Hedlund & Rose, 2020). Transcribers were students who had completed coursework in phonetic transcription. They were blinded to the identity of the child, including the child's age and diagnostic classification. Two students independently transcribed each speech sample, using the blinded transcription function in Phon so they could not see each other's work. Using the blinded consensus function in Phon, a third trained student rater compared the two ratings and resolved any cases of discrepancy by selecting what they judged to be the most appropriate transcription, a procedure akin to the consensus procedure described in McAllister Byun and Rose (2016). Using the automated inter-

transcriber reliability feature in Phon, consonant transcription reliability between the two transcribers was calculated for each child; across all files, the inter-transcriber reliability was 89.15% (where reliability was based on broad transcription with the exception of the diacritics considered distortions, as described below). Transcriptions were then converted into ratings of perceptual accuracy based on whether the transcription of the target phoneme matched the adult target ("correct") or represented a substitution or deletion ("incorrect"). Transcriptions with diacritics indicating prolongation ([æː], [lː], [tː]), contextually appropriate aspiration ([kʰ]), and no audible release ([t̚]) were coded as "correct," while transcriptions with diacritics that did not represent typical allophonic variation (e.g., partial voicing of initial [k̬]) were coded as an intermediate category, "distortion." Partial derhotacization of the /ɹ/ sound ([ɻ]) was coded in the "distortion" category. *Table 3.2* provides a complete confusion matrix accounting for all 1092 transcriptions across all target phonemes, organized by target and transcription. Targets are listed as the row headings with total number of elicitations of each target listed in the rightmost final column. Transcriptions are listed as the column headings with totals of each transcription in the final row. Bolded values are considered "correct," italicized values are considered "distortion," and unmarked values are considered "incorrect." An overwhelming majority of errors in the sample ("incorrect" or "distortion" categories) involved the liquid sounds /l/ and /ɹ/; this was true independent of the disorder status of the speaker. *Table 3.1* reports percentage accuracy in production of /ɹ/ and /l/ for all speakers. Given the prevalence of /ɹ/ and /l/ errors, only these targets were used as

the basis for our third question asking whether tongue complexity differs between correct and incorrect productions in both TD and SSD groups.

Target	ø	æ	ɑ	ɪ	ʌ	j	w	k	ɡ̊	g	t	t̪	t̪̊	ʧ	d	n	f	h	s	ʃ	ʰl	l	ɹ	ɹ̊	Total
æ	-	**214**	2	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	218
j	1	-	-	-	-	**45**	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	3	-	-	50
w	-	-	-	-	-	-	**112**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	112
k	-	-	-	-	-	-	-	**177**	*1*	12	3	-	*1*	-	-	-	-	1	-	-	-	-	-	-	195
t	-	-	-	-	-	-	-	1	-	1	**167**	1	-	2	4	-	-	-	1	1	-	-	-	-	178
l	-	-	-	-	-	6	21	-	-	-	-	-	-	-	-	1	1	-	-	-	*1*	**85**	1	-	116
ɹ	-	-	-	-	-	1	91	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	**94**	*36*	223
Total	1	214	2	1	1	52	224	178	1	13	170	1	1	2	4	1	1	2	1	1	1	89	95	36	1092

Table 3.2. Confusion matrix of all transcriptions.

Analyses

MCI analysis was performed on all individual target sound files using the script 'tshape_analysis' (Dawson, 2016) in Python (Python Software Foundation, 2016). NINFL analysis was performed in Matlab with a custom script, "ComputeCurvature," which is available as supplemental material. Prior to computing the number of inflections (NINFL), the signed curvature was first low-pass filtered (to avoid distortions from localized differences between neighboring contour points), and then thresholded, such that only changes in curvature exceeding the threshold were counted as viable inflections. The filter cutoff (0.075) and threshold (0.002) values were chosen heuristically through a process of visual inspection of representative contours such that NINFL results best accorded with human characterization of the number of distinct tongue shapes observed across the dataset. All subsequent analyses were performed in R (R Core Team, 2019) using RStudio (RStudio Team, 2019). To test the reliability of our primary outcome measures, 20% of all files (n = 5) were retagged on the

previously identified target frame by a second trained tagger. The decision to compare the same frame was based on the precedent set in Preston et al. (2019). For MCI, the intraclass coefficient (ICC) with single random raters was calculated to be 0.56, indicating moderate agreement (Koo & Li, 2016); for NINFL, Cohen's kappa was calculated to be 0.30 which is considered fair agreement (McHugh, 2012). As an additional means of reliability, 75% of all files (n = 18) were tagged by the first author using GetContours' newly-integrated contour tracking feature implementing the SLURP algorithm (Laporte & Ménard, 2018). To ensure that the automated tracking algorithm appropriately tracked all contours, the first author viewed all tagged target frames and made adjustments to the minority of files that were not tracked completely or correctly. Coordinates from the same frames as the original data set were used to calculate MCI and NINFL; reliability across the original and automatically tracked data sets was moderate for both MCI (ICC = 0.62) and NINFL (Cohen's kappa = 0.46).

Following the standards described in Harel and McAllister (2019), all models analyzing MCI used linear mixed-effects regression with the 'lme4' package (Bates, Maechler, Bolker, & Walker, 2015), whereas all models analyzing NINFL data used ordinal mixed-effects regression with the 'clmm' package (Christensen, 2015). Complete data and code to reproduce all figures and analyses in the paper can be retrieved at https://osf.io/3ZHCU/. Complete model outputs can be found in the supplemental materials.

To address our first question of whether there is an association between age and tongue complexity in TD children, we fit two models with MCI (linear

mixed-effects regression) or NINFL (ordinal mixed-effects regression) as the outcome variable and Age and Target (phonemes /j/, /w/, /k/, /t/, /l/, /ɹ/, and /æ/), as well as the Age-Target interaction, as predictor variables. These models also included random intercepts for Word and Child and a by-Child random slope on Target.

Our second question asked whether TD children ages 4-6, who may misarticulate some later developing speech sounds as part of typical development, use more complex contours than peers with SSD. To address this question, we fit two models with MCI (linear mixed-effects regression) or NINFL (ordinal mixed-effects regression) as the outcome variable and fixed effects of Classification (TD versus SSD), Target (phonemes listed above), and the Classification-Target interaction. As above, the models also included random intercepts for Word and Child and a by-Child random slope on Target.

To address the third question of whether TD and SSD children's perceptually incorrect /ɹ/ and /l/ productions show less complex tongue contours than perceptually correct productions of the same phonemes, we fit two models with MCI (linear mixed-effects regression) or NINFL (ordinal mixed-effects regression) as the outcome variable and fixed effects of Accuracy (binary categorical variable with levels "correct" and "incorrect," where distortions as well as substitutions were coded as "incorrect"), Classification (TD versus SSD), and Target (phonemes /ɹ/ and /l/), as well as the two-way and three-way interactions between these factors. As above, these models included random intercepts for Word and Child and a by-Child random slope on Target.

Results

1) Is there an association between age and tongue complexity in TD children?

The models used to address this question involved predicting tongue complexity (MCI or NINFL) from Age (in months, centered around mean age 5;1), Target (7 levels, dummy-coded), their interaction, and random intercepts for Word and Child with a by-Child slope for Target. In both models, the reference Target was /æ/, so all comparisons were made in relation to this level. See model output for MCI on the left of *Table 3.3* and for NINFL on the right of *Table 3.3*. The top section shows that, for the reference Target /æ/, there was no association between age and either measure of tongue complexity. The third section shows

Table 3.3. Output for model predicting MCI (left) and NINFL (right) from Age and Target and the interaction between these two predictors.

	Linear Model Predicting MCI	Ordinal Model Predicting NINFL	
Intercept (/æ/ age 5;1)	**β = 2.44, SE = 0.06, p < 0.0001***		
Age in months (slope for /æ/)	β = -0.007, SE = 0.0084, p = 0.45	β = -0.004, SE = 0.050, p = 0.93	
Target (/w/ vs /æ/)	**β = 0.22, SE = 0.089, p = 0.022***	β = 0.022, SE = 0.38, p = 0.95	
Target (/t/ vs /æ/)	β = -0.062, SE = 0.069, p = 0.38	β = 0.12, SE = 0.28, p = 0.68	
Target (/k/ vs /æ/)	**β = 0.30, SE = 0.067, p = 0.0001***	β = 0.30, SE = 0.33, p = 0.36	
Target (/j/ vs /æ/)	β = 0.092, SE = 0.096, p = 0.34	β = 0.09, SE = 0.44, p = 0.84	
Target (/l/ vs /æ/)	β = 0.17, SE = 0.11, p = 0.13	**β = 0.88, SE = 0.43, p = 0.038***	
Target (/ɹ/ vs /æ/)	**β = 0.43, SE = 0.093, p = 0.0003***	**β = 1.27, SE = 0.41, p = 0.002***	
Age (slope for /w/ vs /æ/)	β = 0.0051, SE = 0.011, p = 0.66	β = 0.034, SE = 0.048, p = 0.48	
Age (slope for /t/ vs /æ/)	β = -0.0049, SE = 0.0088, p = 0.58	**β = -0.078, SE = 0.036, p = 0.030***	
Age (slope for /k/ vs /æ/)	β = 0.0056, SE = 0.0087, p = 0.53	β = -0.025, SE = 0.043, p = 0.56	
Age (slope for /j/ vs /æ/)	β = 0.0079, SE = 0.013, p = 0.55	β = -0.034, SE = 0.057, p = 0.55	
Age (slope for /l/ vs /æ/)	β = 0.0034, SE = 0.014, p = 0.80	β = -0.028, SE = 0.052, p = 0.59	
Age (slope for /ɹ/ vs /æ/)	β = 0.015, SE = 0.012, p = 0.21	β = 0.034, SE = 0.051, p = 0.50	
Random intercept subject	Variance = 0.047, SD = 0.22	Variance = 2.13, SD = 1.46	
Random intercept word	Variance < 0.0001, SD = 0.0003	Variance < 0.0001, SD < 0.0001	
Threshold coefficients:		1|2	β = -0.42, SE = 0.40
		2|3	β = 1.07, SE = 0.41
		3|4	β = 2.99, SE = 0.42
		4|5	β = 4.68, SE = 0.47

that there was no association between age and MCI for any other level of Target; however, there was a significant association between age and NINFL for /t/. The second section shows the differences seen across targets, indicating that the tongue complexity means for /w/, /k/, and /ɹ/ were different from /æ/ based on MCI, and the means for /l/ and /ɹ/ were different from /æ/ based on NINFL. These target-level differences in tongue complexity based on NINFL are visible in *Figure 3.3* (right panel), which also depicts the significant negative association between /t/ and age for NINFL, and the lack of observed differences across age for other phonemes. MCI values are also depicted in *Figure 3.3* (left panel). *Figure 3.4* shows sample perceptually correct /t/ contours from one TD participant who was relatively young (17M, age 4;6) and produced this target with a high NINFL value, and another TD participant who was relatively old (10F, age 6;0) and produced this target with a low NINFL value.

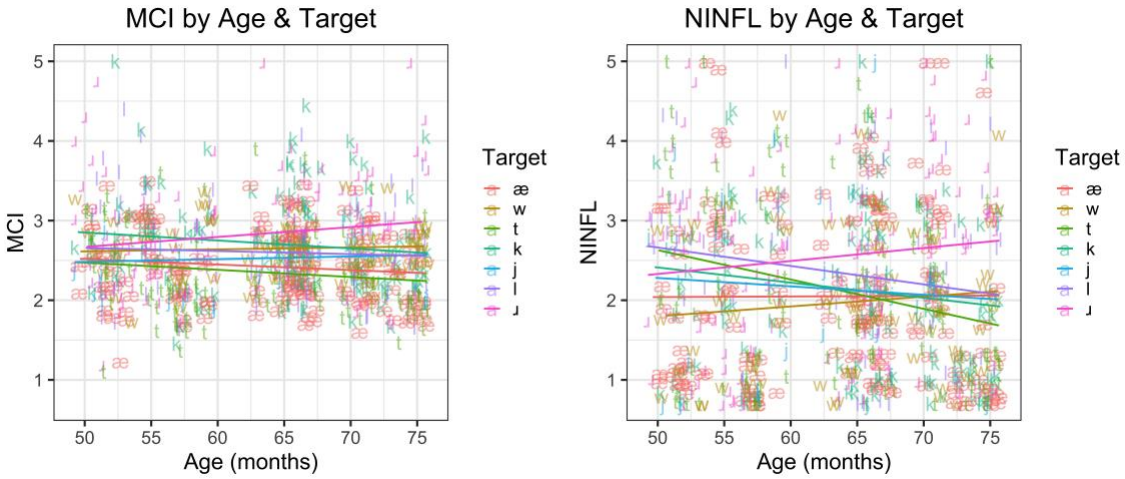

Figure 3.3. Target-level relationship between age (in months) and tongue complexity based on MCI and NINFL.

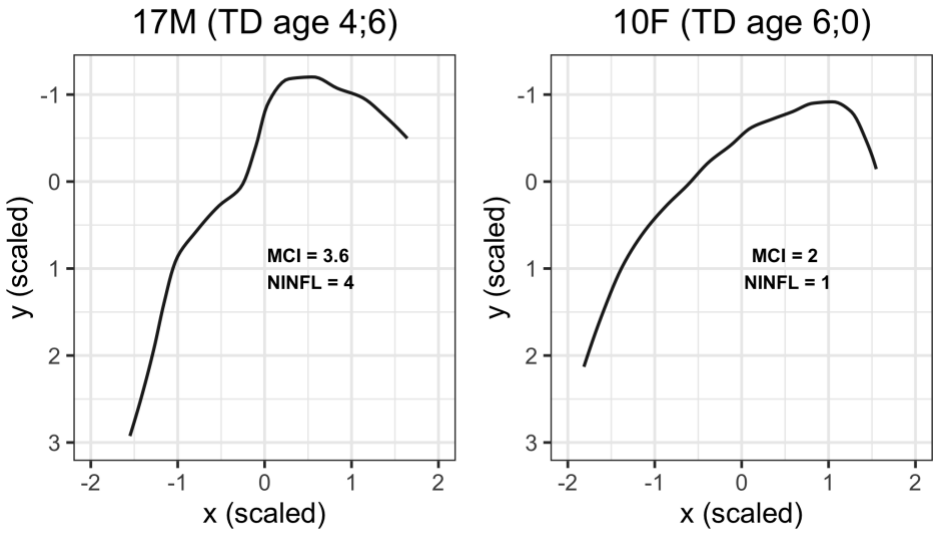

Figure 3.4. Sample /t/ contours from younger and older TD participants depicting high and low NINFL values.

2) Do TD and SSD productions differ in tongue complexity?

The models used to address this question involved predicting tongue complexity from Classification (TD, SSD), Target (7 levels, as above), their interaction, and random intercepts for Word and Child with a random by-Child slope for Target. As before, the reference level for Target was /æ/ for both models. *Table 3.4* presents outputs for the model predicting MCI (left side) and NINFL (right side). The top section shows that there was no significant difference in tongue complexity between children classified in TD versus SSD groups for the reference level /æ/. As before, the second section shows the differences seen across targets. The addition of children with SSD to the model did not change the observed effects of Target: again, the means for /w/, /k/, and /ɹ/ were found to be different from /æ/ based on MCI, while the mean for /ɹ/ was different from /æ/ based on NINFL. The third section refers to the interaction between Classification and Target and shows that the mean complexity for /ɹ/ was significantly greater in

TD children than for SSD children based on NINFL, but not based on MCI.

Figure 3.5 plots MCI against NINFL (jittered) with faceting by Target (7 levels) and color and shape representing Classification (TD, SSD). The targets are ordered by age of mastery according to Smit, Hand, Freilinger, Bernthal, and Bird (1990). The figure shows extensive overlap between data points representing SSD (green triangles) and TD (purple circles) in most contexts, with the greatest separation occurring on the y-axis (NINFL) for /ɹ/.

Table 3.4. Output for model predicting MCI (left) and NINFL (right) from classification (TD, SSD), target, and the interaction between these two predictors.

	Linear Mixed Effects Model Predicting MCI	Ordinal Mixed Effects Model Predicting NINFL	
Intercept (/æ/ TD)	**β = 2.44, SE = 0.065, p < 0.0001***		
Classification (SSD vs TD)	β = 0.060, SE = 0.116, p = 0.61	β = 0.63, SE = 0.69, p = 0.36	
Target (/w/ vs /æ/)	**β = 0.23, SE = 0.10, p = 0.033***	β = 0.059, SE = 0.41, p = 0.89	
Target (/t/ vs /æ/)	β = -0.052, SE = 0.074, p = 0.49	β = 0.20, SE = 0.34, p = 0.56	
Target (/k/ vs /æ/)	**β = 0.31, SE = 0.081, p = 0.0008***	β = 0.34, SE = 0.35, p = 0.32	
Target (/j/ vs /æ/)	β = 0.066, SE = 0.096, p = 0.49	β = -0.01, SE = 0.45, p = 0.98	
Target (/l/ vs /æ/)	β = 0.16, SE = 0.111, p = 0.16	**β = 0.98, SE = 0.43, p = 0.023***	
Target (/ɹ/ vs /æ/)	**β = 0.43, SE = 0.094, p = 0.0001***	**β = 1.27, SE = 0.42, p = 0.0026***	
Classification (SSD vs TD) & Target (/w/ vs /æ/)	β = 0.022, SE = 0.18, p = 0.91	β = -0.22, SE = 0.75, p = 0.77	
Classification (SSD vs TD) & Target (/t/ vs /æ/)	β = -0.095, SE = 0.13, p = 0.47	β = -0.69, SE = 0.61, p = 0.26	
Classification (SSD vs TD) & Target (/k/ vs /æ/)	β = -0.28, SE = 0.15, p = 0.067	β = -0.88, SE = 0.64, p = 0.17	
Classification (SSD vs TD) & Target (/j/ vs /æ/)	β = 0.044, SE = 0.19, p = 0.82	β = -0.43, SE = 0.84, p = 0.61	
Classification (SSD vs TD) & Target (/l/ vs /æ/)	β = -0.050, SE = 0.20, p = 0.81	β = -1.05, SE = 0.79, p = 0.18	
Classification (SSD vs TD) & Target (/ɹ/ vs /æ/)	β = -0.17, SE = 0.17, p = 0.33	**β = -1.61, SE = 0.77, p = 0.035***	
Random intercept subject	Variance = 0.037, SD = 0.194	Variance = 1.81, SD = 1.34	
Random intercept word	Variance = 0.0015, SD = 0.039	Variance < 0.0001, SD < 0.0001	
		Threshold coefficients: 1|2	β = -0.53, SE = 0.38
		2|3	β = 1.17, SE = 0.38
		3|4	β = 3.10, SE = 0.39
		4|5	β = 4.81, SE = 0.43

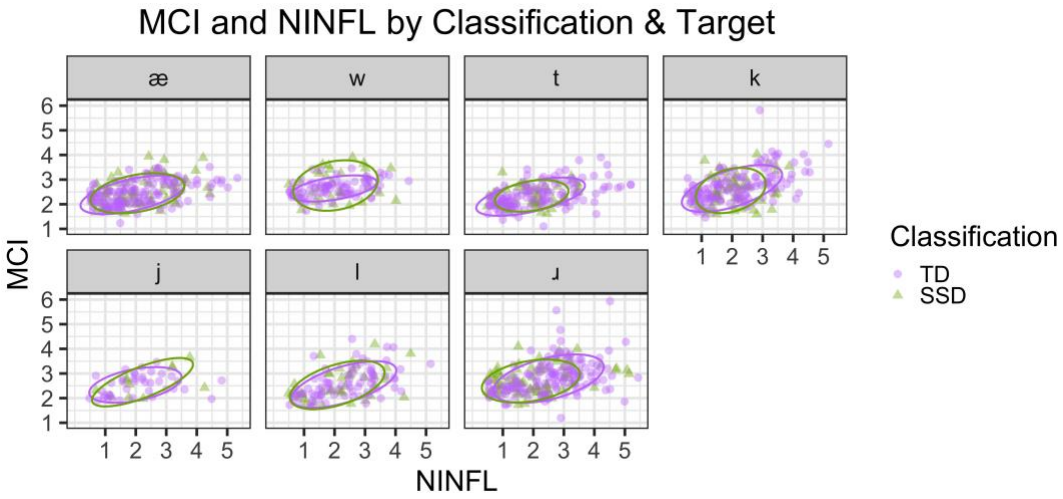

Figure 3.5. Individual MCI and NINFL values separated by target, colored by classification, and ordered by age of mastery.

3) Is there a difference in tongue complexity between /l/ and /ɹ/ productions that differ in perceptually rated accuracy, pooled across TD and SSD groups?

Because /ɹ/ and /l/ were the only phonemes produced with varying accuracy by both groups of children (and also were the only phonemes for which we expected differences in tongue complexity), the models used to address this question included only these phonemes. The models predicted tongue complexity (MCI or NINFL) from perceptually rated Accuracy, Classification, Target (/ɹ/ and /l/ only), their interactions, a random intercept for Word, and a random by-Child slope for Target. In the model examining MCI, neither Accuracy, Classification, Target, nor their interactions were significant predictors. In the model predicting NINFL, there was a significant interaction between Accuracy and Target, such that /ɹ/ tokens rated as perceptually correct had higher tongue complexity than those rated perceptually incorrect, but this was not the case for /l/. See *Table 3.5* for the output of the model predicting MCI (left) and NINFL (right). See *Figure 3.6* (left panel) to visualize the null effect for /l/ and /ɹ/ based on MCI and *Figure*

3.6 (right panel) to visualize the significant interaction between Accuracy and Target based on NINFL in both groups of children. The result based on NINFL suggests that correct /ɹ/ productions have greater tongue complexity than incorrect productions in both TD children and children with SSD.

Table 3.5. Output for model predicting MCI and NINFL from accuracy (correct, incorrect), classification (TD, SSD), target (/l/, /ɹ/), and the interaction between these three predictors.

	Linear Mixed Effects Model Predicting MCI	Ordinal Mixed Effects Model Predicting NINFL	
Intercept (/l/ TD incorrect)	**β = 2.74, SE = 0.18, p < 0.0001***		
Accuracy (correct vs incorrect)	β = -0.16, SE = 0.19, p = 0.41	β = -0.43, SE = 0.61, p = 0.48	
Classification (SSD vs TD)	β = -0.056, SE = 0.25, p = 0.82	β = -0.43, SE = 0.82, p = 0.60	
Target (/ɹ/ vs /l/)	β = 0.021, SE = 0.20, p = 0.92	β = -0.62, SE = 0.59, p = 0.30	
Accuracy (correct vs incorrect) & Classification (SSD vs TD)	β = -0.027, SE = 0.29, p = 0.93	β = 0.17, SE = 0.97, p = 0.86	
Accuracy (correct vs incorrect) & Target (/ɹ/ vs /l/)	β = 0.38, SE = 0.23, p = 0.10	**β = 1.86, SE = 0.70, p = 0.008***	
Classification (SSD vs TD) & Target (/ɹ/ vs /l/	β = 0.094, SE = 0.28, p = 0.74	β = -0.31, SE = 0.87, p = 0.72	
Accuracy (correct vs incorrect) & Classification (SSD vs TD) & Target (/ɹ/ vs /l/)	β = -0.30, SE = 0.36, p = 0.40	β = -0.64, SE = 1.14, p = 0.57	
Random intercept subject	Variance = 0.048, SD = 0.22	Variance = 0.45, SD = 0.68	
Random intercept word	Variance < 0.0001, SD < 0.0001	Variance < 0.0001, SD < 0.0001	
		Threshold coefficients:	1|2 β = -1.60, SE = 0.58
			2|3 β = -0.24, SE = 0.58
			3|4 β = 1.71, SE = 0.59
			4|5 β = 3.33, SE = 0.65

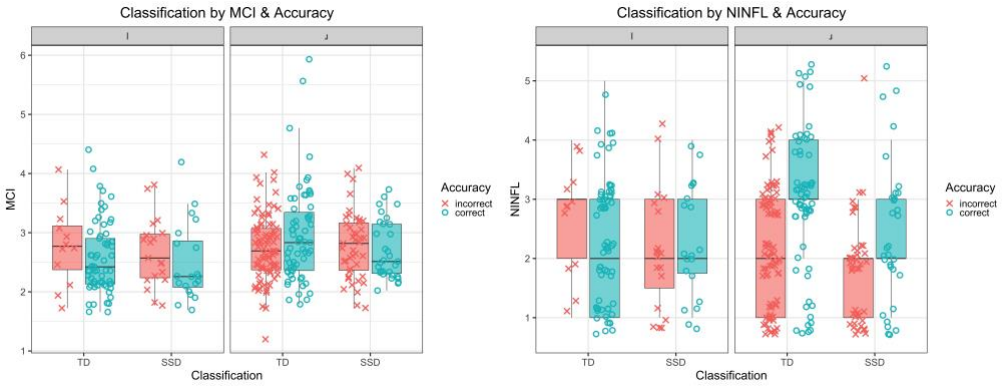

Figure 3.6. Boxplots of MCI and NINFL, faceted by target (/l/ and /ɹ/), separated classification, and colored by binary perceptual rating of accuracy.

Discussion

The present study explored the utility of ultrasound-derived measures of tongue complexity to differentiate between speech sounds produced by American English-speaking children with and without SSD. Our first question asked whether tongue contours would increase in complexity as a function of age within the present sample of TD children (age range 4;0-6;3). While no main effect of age was found, there was a significant interaction in which age was negatively associated with tongue complexity for /t/ as compared to /æ/ for NINFL. In addition, differences in complexity were found between phonemes: /ɹ/ was found to be significantly more complex than /æ/ based on both MCI and NINFL, while /w/, /k/, and /l/ were significantly different from /æ/ based on one of the two measures. Our second question asked whether tongue contours of TD children would be found to be more complex than those of children with SSD. A significant group by target interaction indicated that such a difference existed for the phoneme /ɹ/ as measured with NINFL. The same phoneme-specific differences in tongue complexity were observed in our second question as in the first question. Our final question asked whether perceptually correct and incorrect productions of the phonemes /l/ and /ɹ/ would differ in tongue complexity. Data were pooled across groups for this question, since many of the TD children still produced developmental misarticulations of these late-emerging sounds. The results revealed that correct versus incorrect /ɹ/ productions differed with regard to NINFL but not MCI; no differences were found for correct versus incorrect /l/ productions. Overall, the results of the present study suggest that differences in

tongue complexity can be found across multiple developmental dimensions, supporting the potential of these measures as a correlate of motor control.

Comparison of MCI versus NINFL

Given that MCI and NINFL are both intended to measure tongue complexity, it is important to consider the substantive differences that emerged in the results based on these two metrics. First, in our two initial questions, MCI identified /w/, /k/, and /ɹ/ as different from /æ/, whereas NINFL identified /l/ and /ɹ/ as different from /æ/. Our first question also identified the relationship with age and tongue complexity for /t/ based on NINFL only. Next, for our second question, NINFL identified a greater mean for /ɹ/ in TD children than in children with SSD, but MCI did not identify any phoneme that differed between the two groups. Finally, the third question identified a significant difference between /ɹ/ productions rated as perceptually correct versus incorrect based on NINFL, but not based on MCI. Although there was a moderately strong correlation between MCI and NINFL in the present data set ($r(1090) = 0.46$, $p < .0001$), the discrepancies between the two metrics highlight their computational differences. Because MCI is driven by curvature, it is possible to have a high MCI value with only one point of inflection as long as the local curvature is high (i.e., the radius of curvature is small) at some point, such as near the tongue apex. By contrast, tongue contours only receive high NINFL values if regions of constriction are separated by anti-constriction. Given the complementary nature of the two measures, future research could evaluate whether a metric that combines MCI and NINFL could more fully capture tongue complexity.

In all three cases of discrepancy stated above, the findings from the model examining NINFL align better with theory-driven expectations. That is, given that /l/ and /ɹ/ are late-developing and canonically produced with multiple lingual constrictions by adults, these phonemes are expected to be produced with more complex tongue shapes than other targets. Conversely, /w/ and /k/ are not late-developing or articulatorily complex, so these phonemes are expected to be produced with relatively less complex tongue shapes. Second, previous work such as Gibbon (1999) leads us to expect differences in tongue complexity between TD children and children with SSD, as we observed for NINFL but not for MCI. Finally, perceptually incorrect productions of complex targets such as /l/ and /ɹ/ are expected to involve articulatory simplifications of the tongue shapes used in correct productions (Gick et al., 2007). Preston et al. (2019) observed that NINFL was more successful than MCI in distinguishing tongue shapes for perceptually inaccurate rhotics produced by school-aged children with SSD from perceptually accurate rhotics produced by same-aged children without SSD. Our results extend this finding to younger children. Overall, the results of the present study suggest that NINFL is better equipped than MCI to serve as the index of complexity that represents tongue shape across phoneme targets and across children ranging in vocal tract size and shape. Therefore, the remainder of our discussion will be based on results from NINFL only.

Although we selected NINFL as the best-performing measure, we also acknowledge that it has clear limitations. First, the significant association between age and tongue shape complexity in the TD children was negative, which is

inconsistent with our expectation that tongue complexity should increase over the course of development; we discuss this finding in more detail below. Second, reliability was lower for NINFL than for MCI, particularly in the comparison of two contours that were manually measured by two individuals. Reliability for NINFL improved from fair to moderate when comparing manual and automated measurements, which suggests that automated measurements are a promising direction to investigate in future studies. A final limitation that is true of MCI as well as NINFL is the fact that these metrics only reflect complexity of a single midsagittal cross-section of the tongue. This approach ignores information that could be obtained by imaging the parasagittal tongue, which is relevant for the many tongue shapes that are produced with active lateral bracing (Gick, Allen, Roewer-Després, & Stavness, 2017) and for acoustic shaping generally. Measures based on multiple planes through the tongue or three-dimensional ultrasound (Lulich & Pearson, 2019), together with consideration of how these interact with speaker hard structure morphology (Brunner, Fuchs, & Perrier, 2009), will be necessary to provide a complete representation of tongue shape complexity.

Implications and future directions

In response to our first question, a negative relationship was observed between age and tongue complexity for the target phoneme /t/. It was surprising that /t/ would differ in tongue complexity between the relatively younger and older children in our TD sample because this target is relatively early-developing and was generally transcribed as correct. Furthermore, this pattern was incongruent with our expectation that tongue complexity would increase over the

course of development. That is, the association between age and tongue complexity for /t/ was negative, with younger children tending to show more complex tongue shapes than older speakers. Although we interpret this age-based pattern observed for /t/ with caution, it could reflect a pattern in which some younger TD children produce /t/ with increased tongue complexity as they are exploring different articulatory strategies to achieve the desired phonetic output (Guenther, 2016; Schwartz et al., 2012). A related possibility is that older children may have developed more complex articulatory patterns, but that this complexity would only be apparent in a coronal section, not in the midsagittal plane measured here. This is compatible with previously mentioned research such as Gibbon's (1999) description of a "spoon-like" pattern of contact for differentiated alveolar stops (Gibbon, 1999; Gick et al., 2017); it also speaks to the importance of examining complexity in multiple planes of section. In either case, this finding represents a covert difference between tokens that were transcribed identically. Thus, it is broadly compatible with the findings of covert error in children with SSD that were argued to offer support for the clinical utility of ultrasound imaging in Cleland et al. (2017).

The absence of a general association between age and tongue complexity may be an artifact of the small age range of the children in our data set. The ages of the TD children ranged only from 50 to 75 months, and robust differences in tongue complexity may not be detectable within this relatively narrow window. The comparison was also made cross-sectionally rather than longitudinally, reducing the likelihood of being able to observe strong age-based differences.

Future comparisons should include children representing a wider range of ages, although it is our opinion that collecting ultrasound data from children *younger* than those included in the present study using the current word list and procedures would be extremely challenging. Future analyses might thus compare contours from the children in the present sample with a group of older children producing the same targets in the same task. It would also be ideal to include a longitudinal component, particularly if cross-sectional results can be used to identify a time window likely to represent the most active period of change. It remains unknown at what point in childhood tongue complexity becomes adultlike, and it will be important to understand how changes in tongue complexity relate to the trajectory of perceptual and acoustic measures of speech from childhood to maturity.

In answer to our second question, we found a significant difference in tongue complexity between TD children and children with SSD, such that TD children were found to have more complex tongue contours for /ɹ/. The pattern whereby TD children produced /ɹ/ with more complex contours than children with SSD is congruent with the fact that this target is articulatorily complex, phonologically marked, late-developing, and among the most commonly misarticulated phonemes. Because the interaction with group was only found for /ɹ/, the present results do not provide evidence for global differences in tongue complexity between children with and without SSD. This contrasts with previous studies suggesting that children with SSD may exhibit globally reduced tongue complexity, including covert errors on sounds that are perceived as accurate (Cleland et al., 2017; Gibbon, 1999). However, as noted in the introduction,

participants in Cleland et al. (2017) and Gibbon (1999) were specifically selected for the presence of persistent speech errors with a suspected motoric origin, whereas the younger children in this study are likely to be more heterogeneous in the nature of their SSD. Specifically, it is likely that at least some of our participants' speech deficits were primarily phonological rather than motoric in origin, which is likely to have limited our ability to observe global differences in tongue complexity between TD and SSD groups. Although we did not find a difference in overall tongue complexity between diagnostic groups, the present findings did reinforce the fact that ultrasound can be used to detect differences in tongue complexity that are not reflected in transcription. Thus, our findings are broadly supportive of previous research suggesting that tongue complexity measures could potentially assist with diagnosis and treatment planning for children with suspected SSD. However, there is a need for considerable further research before this clinical potential can be realized. Future work will need more participants in both diagnostic categories, ideally including a longitudinal component. It could be particularly valuable to quantify a baseline level of tongue complexity and track progress over the course of different types of treatment.

In answer to our third question, perceptually accurate /ɹ/ productions from children with and without SSD were found to have more complex tongue contours than the incorrect /ɹ/ productions from these same children. This finding echoes the observation by Klein et al. (2013) that the degree of lingual differentiation of rhotic gestures is associated with degree of perceived accuracy. However, no significant relationship was found between tongue complexity and accuracy for /l/

in our pooled child data set. We had expected a similar association between tongue complexity and perceptual ratings of accuracy for laterals based on evidence that adult speakers of American English produce onset /l/ with two lingual constrictions, and that misarticulation of /l/ may involve gestural simplification of this complex tongue shape (Gick et al., 2007). However, as mentioned previously, Lin and Demuth (2015) found that young children may use a single lingual constriction to produce perceptually accurate onset /l/, and that the use of multiple constrictions for this target increases with age in early childhood. Thus, it is possible that the hypothesized relationship occurs in a later stage of development than that represented in our sample.

Finally, even though the relationship between tongue complexity and accuracy ratings was significant for /ɹ/, it can be observed in *Figure 3.6* that there are many exceptions such as perceptually accurate /ɹ/ productions with low tongue complexity and perceptually incorrect /ɹ/ productions with high tongue complexity. These exceptions further suggest that tongue complexity reveals information that is distinct from perceptual ratings of accuracy. Specifically, the finding that both TD children and children with SSD produced some low complexity contours for targets that usually have articulatorily complex contours (such as /l/ and /ɹ/) can be seen as consistent with the claim that undifferentiated gestures are a pervasive developmental phenomenon, as found in Gibbon (1999). Of course, both perceptual ratings and physiological measurements are subject to noise, so instances in which tongue complexity and perceptual ratings disagree could highlight tokens that could be fruitfully examined in greater detail.

Furthermore, some of the variation in tongue complexity for perceptually accurate /ɹ/ may be attributable to natural variation in tongue shape, particularly between "bunched" versus "retroflex" contours (Delattre & Freeman, 1968; Tiede, Boyce, Holland, & Choe, 2004). However, Stolar and Gick (2013) found that MCI values were similar across retroflex and bunched tongue shapes in their study of typical adult speakers. In their study of NINFL values in children's tongue shapes, Preston et al. (2019) avoided classifying tongue shapes as retroflex versus bunched due to the large number of perceptually accurate rhotics produced with tongue shapes that do not match either of these classical categories. We avoid a systematic comparison of tongue complexity in retroflex versus bunched tongue shapes for the same reason, but future work may wish to address this question with the subset of tongue shapes that conform to the classical retroflex and bunched categories.

Taken together, our results point to the utility of ultrasound-based tongue complexity metrics, particularly NINFL, in revealing information about speech development in young children, beyond what may already be available through perceptual ratings of accuracy. This research highlights the value of ultrasound as an increasingly affordable and child-friendly tool, extending results from earlier research using EPG (Fletcher, 1989; Gibbon, 1999). Our findings also support the idea that ultrasound imaging can be used to identify covert differences in tongue shape (Cleland et al., 2017), since younger and older TD speakers differed in tongue complexity for /t/, even though it was generally transcribed as perceptually accurate. The presence of globally reduced tongue complexity could potentially

be used to support the diagnosis of SSD and may also suggest a disorder that is more motoric than phonological in etiology. Although the immediate clinical applicability of tongue complexity measures remains limited by the cost of the equipment and the time-intensive nature of the analyses, ongoing technological advances can be expected to expand the accessibility of the clinically valuable measures examined here.

Acknowledgments

This research was supported by the National Institute on Deafness and Other Communication Disorders Grant F31DC018197 (H. Kabakoff, PI), Grant R01DC013668 (D. H. Whalen, PI), and Grant R01DC017476 (T. McAllister, PI). Additional support was provided through an Acoustical Society of American Stetson Scholarship and an American Speech-Language-Hearing Foundation New Century Scholars Doctoral Scholarship. The authors gratefully acknowledge Jonathan L. Preston for supervision of the project across multiple sites, Twylah Campbell, Emily Phillips, and Megan Leece for data collection, Megan Leece for coordination of phonetic transcriptions, and Graham Tomkins Feeny for continuous coordination and support measuring and cleaning ultrasound data accumulated across multiple sites. We also acknowledge Kevin Roon for guidance with Praat and Matlab, and Katherine M. Dawson for sharing Python scripts. We thank Siemens Medical Solutions USA, Inc., for making their Acuson ultrasound scanner available for this project.

MANUSCRIPT 3: Characterizing sensorimotor profiles in children with residual speech errors

Abstract

Purpose: Past research suggests that children with speech errors who have reduced motor skill may be more likely to develop residual errors associated with lifelong challenges. Drawing on models of speech production that highlight the role of somatosensory acuity in updating motor plans, this pilot study explored the relationship between motor skill and speech accuracy, and between somatosensory acuity and motor skill in children. Understanding the connections among sensorimotor measures and speech outcomes may offer insight into how somatosensation and motor skill cooperate during speech production, which could offer insight into treatment decisions for this population.

Method: Twenty-five children (ages 9-14) produced syllables in an /ɹ/ stimulability task before and after an ultrasound biofeedback treatment program targeting rhotics. We first tested whether motor skill (as measured by two ultrasound-based metrics of tongue complexity) predicted acoustically measured accuracy. We then tested whether somatosensory acuity (as measured by an oral stereognosis task) predicted motor skill, while controlling for auditory acuity.

Results: One measure of tongue complexity was a significant predictor of accuracy, such that higher tongue complexity was associated with lower accuracy at pre-treatment but higher accuracy at post-treatment. Children with better sensory acuity in either domain produced /ɹ/ tongue shapes that were more complex, but this relationship was only present at post-treatment.

Conclusion: The predicted relationships among somatosensory acuity, motor skill, and acoustically measured /ɹ/ production accuracy were observed after treatment, but unexpectedly did not hold before treatment. This finding that greater degrees of somatosensory or auditory acuity are associated with more complex tongue shapes and greater production accuracy after participation in ultrasound biofeedback treatment has the potential to inform future research seeking to match individual children to treatment approaches tailored to their specific area of deficit.

Introduction

Clinical background

Individuals differ in their level of skill in various sensorimotor domains, and differences among these skills may be associated with variance in speech production outcomes. Of particular interest to the present study are the 25% of children with speech sound disorder (SSD) that show persisting errors past age six (Shriberg et al., 1999), as well as the 1-2% of the population that persist with residual speech errors (RSE) into adolescence/adulthood (Flipsen, 2015). RSE can lead to lifelong challenges, so knowing which sensorimotor factors predict who will persist with errors beyond childhood is a crucial step toward making evidence-based clinical decisions for this population. According to the Speech Disorders Classification System, SSD has three typological branches: speech delay; speech errors; and motor speech disorder (Shriberg et al., 2010). Children with motor involvement are considered most likely to develop persistent errors

(Vick et al., 2014; Wren, Miller, Peters, Emond, & Roulstone, 2016), but the means available for measuring motor involvement are limited.

English /ɹ/ is among the most motorically challenging speech sounds, which partially explains why it is considered one of the most common residual errors (Ruscello, 1995). This articulatory complexity is characterized by the combination of posterior and anterior lingual constrictions with lateral lingual bracing and lip rounding. Perceptually accurate /ɹ/ can be achieved using a variety of tongue shapes, such as retroflex and bunched configurations (Delattre & Freeman, 1968; Tiede et al., 2004). However, the articulatory shaping strategy that will work best for a given individual cannot readily be determined, so treatment for this target can be challenging and long-lasting (Ruscello, 1995). For these reasons, the present study will focus on RSE affecting rhotic targets in American English.

Sensorimotor integration in speech production

According to current models of speech production, speech is produced by executing stored motor plans to achieve targets in auditory and somatosensory space (e.g., Guenther, 2016; Hickok, 2012; Houde & Nagarajan, 2011). Over the course of typical development, children undergo a process of attempting various gestures and configurations until they arrive at motor plans that map onto auditory-perceptual goals (Guenther, 2016; Schwartz et al., 2012). As children work toward the motor plan that best helps them achieve auditory speech targets, they gradually form somatosensory targets corresponding with their tactile and proprioceptive experience of the production of various articulatory gestures. Over

time, precision develops in the specification of both auditory and somatosensory targets, which increases the efficiency of the feedforward control system in achieving sensory targets in a range of coarticulatory contexts (e.g., Guenther, 2016). We refer to this emergent robustness of the feedforward plan as "motor skill."

Execution of motor plans during speech production is modulated by auditory and somatosensory feedback (e.g., Guenther, 2016; Hickok, 2012; Houde & Nagarajan, 2011). The ability to use feedback in each sensory domain is related to the ability to detect fine-grained detail or classify stimuli presented in that domain; we refer to these skills as "auditory acuity" and "somatosensory acuity." Here we conceptualize motor skill, auditory acuity, and somatosensory acuity as distinct but interacting sensorimotor factors that influence a speaker's ability to execute stored motor plans and access and respond to sensory feedback in order to update those motor plans (Tremblay et al., 2003). In order to provide a complete characterization of children's sensorimotor profiles as they relate to speech production skill, the current study considers all three skills in connection with acoustically measured production accuracy. Specifically, the objective of this pilot study is to examine the relationship between speech production skill and motor and sensory factors in children with RSE affecting rhotic targets.

Tongue complexity as an index of motor skill

As suggested above, some target phonemes naturally require a greater degree of motor skill than other targets, and these relatively more motorically complex articulatory targets are generally acquired later than those that are

produced with simpler articulatory configurations. Studdert-Kennedy and Goldstein (2003) described four stages of articulatory development for consonants in English, where the first stage involves acquisition of voiceless stops, nasals, and glides, the second stage involves acquisition of the laryngeal control needed to produce voicing contrasts, the third stage involves refinement of control of jaw height and lingual constriction needed to produce fricatives and affricates, and the final stage involves acquisition of multiple lingual gestures needed to produce laterals and rhotics. This trajectory of articulatory development is broadly compatible with developmental stages of phoneme acquisition that are based on perceptually-determined accuracy (Crowe & McLeod, 2020; Shriberg, 1993; Smit et al., 1990).

Previous research has quantified motor skill in children by measuring the extent to which articulators operate synergistically during speech production (e.g., Green, Moore, Higashikawa, & Steeve, 2000; Noiray, Abakarova, Rubertus, Krüger, & Tiede, 2018) and the degree of movement consistency of the articulators (Goffman & Smith, 1999; Grigos, 2009). These previous studies have used electro-magnetic articulography (EMA) and other kinematic measures that focus on the lips, jaw, and tongue, and are limited to anterior speech targets. An additional line of research has measured patterns of tongue-palate contact using static palatography or electropalatography (EPG) and has found that the ability to move anterior versus posterior regions of the tongue semi-independently increases over the course of development (Fletcher, 1989). Case studies using EPG have suggested that reduced capacity for isolated movement of lingual regions (i.e.,

reduced *lingual differentiation*, a concept we expand on below) occurs more often in individuals with atypical speech than in children with typical speech (Gibbon, 1999). While EMA- and EPG-based methods have been useful to measure motor skill for a subset of relatively early-developing speech targets (e.g., bilabials with EMA; coronal/dorsal stops with EPG), the complex tongue shapes that characterize rhotic targets are not fully visible using EMA or EPG. Thus, the present study used ultrasound as a means of visualizing the continuous midsagittal tongue shape in order to quantify motor skill for rhotic targets.

Lingual differentiation refers to an individual's degree of separable control over anterior versus posterior lingual regions. As shown by the taxonomy of articulatory development described above, increased motor control over the tongue is associated with a greater degree of lingual differentiation; conversely, reduced motor control over the tongue is associated with a lower degree of lingual differentiation. This claim is further substantiated by findings that lingual differentiation is connected with achievement of adultlike speech in typically developing (TD) populations (Abakarova, Iskarous, & Noiray, 2020; Fletcher, 1989). For lateral targets that require simultaneous anterior and posterior lingual constrictions, TD children ages 3-7 tend to simplify the articulation to involve one lingual constriction (Lin & Demuth, 2015). For rhotic targets, children's error patterns involve motor simplifications characterized by elimination of either the anterior or posterior lingual constriction, or joining of the two constrictions into one central undifferentiated lingual constriction (Gick et al., 2007). These articulatory simplification patterns in typical development suggest that measuring

degree of lingual differentiation over the course of development could be helpful in determining an individual's degree of motor skill. Lingual differentiation has also been found to differ between TD speakers and children with SSD (Gibbon, 1999; Green et al., 2000). Gibbon (1999) found that the majority of children ages 4-12 with SSD showed pervasive undifferentiated gestures that involved broad linguopalatal contact exceeding what is typical in adult production. In an articulatory study with children aged 5-6 with SSD affecting rhotic targets, Klein et al. (2013) used ultrasound to test for associations between midsagittal tongue shapes and perceived accuracy. They found that more differentiated lingual gestures were related with higher degrees of perceived accuracy, and that lingual differentiation increased over the course of treatment targeting rhotics. This finding suggests that children with delays in motor control may be able to achieve the required degree of lingual differentiation with the help of treatment, which will be discussed in a later section.

Recent research suggests that ultrasound-based measures can be a valid means to evaluate the degree of lingual differentiation of a given midsagittal tongue shape or 'contour.' Methods of tongue measurement in previous ultrasound research include the use of ratio-based measures for quantifying positional attributes of individual productions (Klein et al., 2013; Ménard et al., 2012; Zharkova et al., 2015). While this approach is helpful for describing the position and shape of contours with one lingual constriction, it is not suitable for describing contours with multiple lingual constrictions, such as /l/ and /ɹ/. Instead of using these previously established approaches to the analysis of tongue shapes

using ultrasound data, the current study adopted metrics based on degree of curvature or "tongue shape complexity" because such measures are able to represent contours with multiple constrictions. Furthermore, such metrics are considered robust to differences in translation, scaling, and rotation, which is a prerequisite for valid comparisons within and between speakers who differ in vocal tract size, and under different imaging conditions (Ménard et al., 2012; Stone, 2005).

Previous research comparing several tongue complexity measures (Kabakoff et al., under review) found that a modified curvature index (MCI) and the Number of INFLection points (NINFL) were appropriate for the measurement of lingual contours produced by child speakers. MCI (Dawson et al., 2016) is the integral of absolute curvature (reciprocal of the tangent circle) at each point along a given contour. For adults, Dawson et al. (2016) reported higher MCI values in phonemes with multiple lingual constrictions (/ɹ/ and /l/) than in contours with a single lingual constriction (/æ/ and /ɪ/). For young children age 4-6, higher MCI values were found in /w/, /k/, and /ɹ/ than in a vowel tongue shape selected to represent a low degree of complexity (Kabakoff et al., 2021). NINFL (Preston et al., 2019) represents the number of sign changes in curvature along a given contour, with pre-set thresholding to discard trivial local fluctuations in curvature. For school-aged children producing /ɹ/, higher NINFL values were found in TD children relative to children with RSE, in correct productions relative to incorrect productions, and at post-treatment relative to pre-treatment (Preston et al., 2019). Similarly, for younger children, Kabakoff et al. (2021) reported higher NINFL

values for /ɹ/ and /l/ than for a reference level /æ/ shape, higher NINFL values for /ɹ/ in TD children relative to children with SSD, and higher NINFL values for /ɹ/ in perceptually correct relative to incorrect productions. Building on this recent set of promising results, the current study used MCI and NINFL to quantify degree of lingual differentiation as an index of motor skill in children with RSE.

Auditory acuity

Of the two sensory skills most relevant to speech production, the auditory domain is the most thoroughly studied; it will therefore be treated as a controlled covariate in the present study, which focuses on the less-studied somatosensory domain. Perceptual discrimination and identification tasks are the most commonly used ways to measure auditory-perceptual skill. In a discrimination task, the listener must identify perceptual stimuli as the same or different; for instance, in an ABX task, the listener must decide whether the third stimulus is the same as the first or the second stimulus. Discrimination tasks provide information about an individual's ability to detect fine-grained acoustic differences between stimuli, including within-category differences; by manipulating the magnitude of the difference, it is possible to arrive at the discrimination threshold for a given listener. In a classic identification task, an individual listens to a series of stimuli (typically a synthetic continuum between two sounds) and performs a forced-choice classification of stimuli into phonemic categories. For analysis, a sigmoid function is fit to the responses to reveal the boundary location, or the acoustic crossover point at which listeners shifted their identification responses from one category to the other. In addition to boundary location, previous studies have

examined the consistency with which listeners classify the stimuli along the continuum as an index of sensitivity to within-category information. This degree of consistency can be quantified as the width of the interval between the 25% and 75% probability points along the sigmoid function (Hazan & Barrett, 2000; McAllister Byun & Tiede, 2017).. Although discrimination thresholds from a discrimination task and category labeling consistency from an identification task represent distinct aspects of auditory perception, some previous literature (e.g., Ghosh et al., 2010; McAllister Byun & Tiede, 2017; Perkell et al., 2004) has applied the general term "auditory acuity" to refer to performance on either of these measures.

Various studies have related auditory acuity with speech production skill. Using discrimination tasks, Perkell et al. (2004) and Ghosh et al. (2010) found that adults who were more able to discern small differences between sibilant contrasts also showed more distinctness in production for the same phonemes. Extending this finding to TD children ages 9-14, McAllister Byun and Tiede (2017) found that children with greater category labeling consistency for a synthetic continuum from *rake* to *wake* also produced the /ɹ/ sound with a greater degree of acoustically measured accuracy (i.e., a smaller difference between the second and third formants). A relationship between auditory acuity and production accuracy was also supported by recent work that used the same identification task as McAllister Byun and Tiede (2017). Preston, Hitchcock, and Leece (2020) found that children's auditory acuity was associated with degree of improvement in an ultrasound biofeedback treatment program, independent of

whether the children received perceptual training. Cialdella et al. (2021) found that TD children had higher auditory acuity than children with RSE. However, Cialdella et al. (2021) did not find any association between auditory-perceptual acuity at baseline and /ɹ/ production accuracy at baseline in their sample of children with RSE. In fact, other factors included in the statistical model such as age and gender and the interaction of these factors with auditory-perceptual acuity were also overall model fit was nonsignificant. This overall nonsignificant model fit suggests that other factors not incorporated into the model (such as somatosensory acuity) could be important for understanding individual differences in /ɹ/ production accuracy. In light of such findings, the present study of /ɹ/ production in children with RSE controls for auditory acuity while focusing on the less studied somatosensory domain.

Somatosensory acuity

Degree of refinement of somatosensory goals and the ability to access and respond to somatosensory feedback have also been shown to be essential in speech motor control (e.g., Hickok, 2012; Tremblay et al., 2003). Research into limb movements (e.g., Berryman, Yau, & Hsiao, 2006) has suggested that somatosensation involves two distinct skills that each contribute separately to the ability to respond to somatosensory feedback. Tactile awareness refers to the ability of mechanoreceptors in the skin to identify when contact is made with an object, whereas proprioceptive awareness refers to a body's ability to identify the position and movements of its parts, derived from muscle spindle fibers as well as mechanoreceptors.

The importance of oral somatosensation for speech production has been documented through various lines of research. Post-lingually deafened individuals are able to maintain the ability to produce intelligible speech for some time (Nasir & Ostry, 2008), suggesting that stored motor plans in combination with intact somatosensory feedback are able to function even in the absence of auditory feedback. Additionally, individuals with temporarily reduced access to somatosensory feedback following oral anesthesia tend to produce articulatory errors characterized by reduced precision in movement (Borden, Harris, & Catena, 1973; Gammon, Smith, Daniloff, & Kim, 1971; Putnam & Ringel, 1976; Ringel & Steer, 1963). Research into the proprioceptive aspect of somatosensory acuity has focused on oral perturbations introduced via a bite block or palatal prosthesis (e.g., Baum & McFarland, 1997; Zandipour et al., 2006) or via mechanical manipulation of the articulators (e.g., Feng, Gracco, & Max, 2011; Lametti, Nasir, & Ostry, 2012). Many of these studies allow for the use of auditory feedback in the presence of the oral perturbation, whereas others have atttempted to isolate the role of somatosensory feedback by blocking auditory feedback with masking noise (Gritsyk et al., 2021; Zandipour et al., 2006), or have introduced perturbations that do not impact the acoustic output (e.g., horizontal jaw displacement; Nasir & Ostry, 2006; Tremblay, Shiller, & Ostry, 2003). The results of these studies indicate that individuals are able to use somatosensory feedback to adapt their articulatory trajectories in compensation for mechanical perturbations, including in the absence of auditory feedback. In an exploration of the connection between sensory acuity and speech production skill,

Ghosh et al. (2010) measured somatosensory and auditory acuity in adults producing /s/ and /ʃ/. They found that those participants with stronger auditory acuity and somatosensory acuity exhibited a relatively greater acoustic distinction between the two sibilants. Taken together, previous research has established that an intact somatosensory feedback loop (in connection with an intact auditory feedback loop) is an essential component for the production of precise speech movements.

Somatosensory acuity has been measured in various ways, most of which assess tactile awareness with reference to an individual's ability to detect pressure or vibration or to identify the form of an object presented in the oral cavity (Attanasio, 1987). In vibrotactile threshold detection tasks, voltages are applied to various articulators (tongue, lips) to determine the minimum amplitude at which vibrations at different frequencies can be detected (Fucci, 1972). Similarly, the pressure at which participants can detect the presence of bendable filaments pressed against the lips and tongue has been measured as an index of tactile awareness (Etter, Miller, & Ballard, 2017). Other tasks tap participants' ability to detect small differences, including a two-point discrimination task measuring how far away two points need to be for an individual to detect the presence of one versus two points (McNutt, 1977) and an oral form discrimination task in which individuals feel two items in the oral cavity and determine whether they are the same or different (Ringel, Burk, & Scott, 1968). Other studies have measured the tactile ability to identify the form of an object, including a grating orientation task in which participants identify the direction of lines on an object in the mouth

(Ghosh et al., 2010). Fucci and Robertson (1971) used an oral stereognosis task in which individuals identified the form of a geometric object using only the tongue; more recently, Steele, Hill, Stokely, and Peladeau-Pigeon (2014) introduced a similar task in which individuals used their tongue tip to identify a letter embossed on a plastic strip.

When selecting a task to measure somatosensory function, it is important to acknowledge that no one task constitutes a pure measure of somatosensory acuity. For example, many tasks measuring the tactile aspect of somatosensory acuity (e.g., grating orientation and oral stereognosis) recruit spatial awareness skills by requiring mental rotation of shapes or letters. Although no research has systematically tested which task may be best suited for measuring somatosensory skill in children, previous research using tactile acuity tasks (an oral stereognosis task and a two-point discrimination task) has revealed lower somatosensory acuity in adolescents with RSE than TD peers (Fucci & Robertson, 1971; McNutt, 1977). Furthermore, analysis of recently collected pilot data suggests that performance on an oral stereognosis task differs between child and adult groups (Kabakoff et al., 2020), which could reflect maturation of somatosensory acuity over the course of development (although maturation of other skills such as spatial awareness cannot be ruled out). In light of these previous findings, as well as considerations of ease of administration with child populations, the present study adopted the letter-based stereognosis task operationalized by Steele et al. (2014) and used in Gritsyk et al. (2021) and Kabakoff et al. (2020).

Connection between tongue complexity and somatosensory acuity

Taking the previously described evidence together, it is reasonable to hypothesize that somatosensory acuity and tongue complexity may be correlated with one another. First, group comparison studies suggest that children with RSE differ from their TD peers with respect to both tongue complexity for later-developing sounds like /ɹ/ (Kabakoff et al., 2021; Preston et al., 2019) and somatosensory acuity (Fucci & Robertson, 1971; McNutt, 1977). Production of accurate /ɹ/ is associated with a particularly complex tongue shape, and speakers may need to engage in extensive exploration of different articulatory-acoustic mappings before arriving at a tongue shape that achieves the desired auditory target. It is reasonable to posit that the robustness of an individual's ability to access and respond to somatosensory feedback could influence their ability to explore and refine tongue shapes to attain the desired acoustic consequence. More directly, if achievement of articulatory targets relies on somatosensory feedback, it follows that somatosensory skill may be a predictor of motor skill (i.e., tongue complexity). However, very little previous work has directly investigated this hypothesized association. The present study tested for such an association in children with RSE affecting /ɹ/.

The current study

The first goal of this study was to quantify the relationship between motor skill, as measured by tongue complexity, and acoustically measured production accuracy of /ɹ/ sounds from children with RSE before and after speech remediation. Based on previous research indicating that differentiated tongue

shapes are required for accurate production of later developing targets such as /ɹ/ (Gick et al., 2007; Kabakoff et al., 2021; Preston et al., 2019), we hypothesized that tongue complexity would be directly associated with acoustically measured production accuracy. The second goal was to determine whether somatosensory acuity, as measured by an oral stereognosis task, is associated with degree of tongue complexity in individuals with RSE. We predicted that children with higher somatosensory acuity would have more complex tongue shapes for the target phoneme /ɹ/, consistent with the above-described hypothesis that higher somatosensory acuity should lead to better use of somatosensory feedback in order to achieve complex articulatory targets. In light of previous evidence documenting links between auditory perception and production accuracy, we also controlled for auditory acuity by including it as a controlled covariate in our statistical models. Understanding the connections between motor skill, somatosensory acuity, auditory acuity, and acoustically measured production accuracy may offer insight into how these domains combine to shape speech outcomes in both typical and clinical populations.

Methods

Participants

Participants were 25 children (ages 9;0-14;7, mean = 10;7, SD = 1;5) with RSE affecting rhotic sounds who completed a standard ten-week course of ultrasound biofeedback treatment at Haskins Laboratories or New York University (NYU). All were native speakers of American English with no history of hearing impairment or neurocognitive disorder, per parent report. Inclusionary

criteria were that participants showed normal structure and function of the oral mechanism, passed a pure-tone hearing screening at 500, 1000, 2000, and 4000 Hz at 20 dB HL, displayed average language skills based on a score at or above 80 on the Peabody Picture Vocabulary Test, Fourth Edition (PPVT-4; Dunn & Dunn, 2007). They also had to score 1 or lower on both dysarthria and apraxia indices of the Maximum Performance Tasks (Rvachew, Hodge, & Ohberg, 2005), suggesting the absence of these conditions. To document the presence of RSE, participants were required to score below the 8th percentile on the Goldman Fristoe Test of Articulation, Second Edition (GFTA-2; Goldman & Fristoe, 2000), and to score below 30% accuracy when producing words from a standard list assessing /ɹ/ accuracy, based on ratings by the treating clinician, a certified speech-language pathologist. Children also completed the Recalling Sentences subtest of the Clinical Evaluation of Language Fundamentals, Fifth Edition (CELF-5 ; Wiig et al., 2013) and the Phonological Awareness subtests of the Comprehensive Test of Phonological Processing-2 (CTOPP-2, Wagner, Torgesen, Rashotte, & Pearson, 2013).

Schedule

Participants came in for two sessions lasting approximately two hours each on separate days in order to complete all evaluative tasks. All inclusionary tasks were administered in the first evaluative session, including the hearing screening, the oral mechanism screening, the PPVT-4, the GFTA-2, the Maximum Performance Tasks, and the administration of four production probes while recording audio only. The four production probes were an /ɹ/ word probe

(50 distinct words), an /ɹ/ stimulability probe described in greater detail below, an /ɹ/ sentence probe (5 sentences), and a probe eliciting varied consonants in word-initial position (15 words elicited three times each). See supplemental materials for a complete list of items elicited in all four tasks. At the second evaluative session, the CTOPP-2 and CELF-5 subtests were administered. Additionally, the oral stereognosis and *rake-wake* perceptual identification tasks were administered (both tasks described in detail below), and the four production probes listed above were re-elicited while recording audio and ultrasound simultaneously.

Once enrolled in the study, participants completed one week of intensive traditional therapy (three 1.5 hour sessions), one week of intensive ultrasound biofeedback therapy (three 1.5 hour sessions), and eight weeks of lower-intensity ultrasound biofeedback therapy (two 1-hour sessions weekly). Each intensive session featured 50 minutes of pre-practice and 30 minutes of structured practice intended to elicit 96 trials; each regular session featured 15 minutes of pre-practice and 45 minutes of structured practice intended to elicit 216 trials. After all treatment was complete, participants came in for a final session to complete the same *rake-wake* perceptual identification task and the same four production tasks with simultaneous audio and ultrasound recording. See *Table 4.1* for scores from evaluative tasks and post-treatment tasks for each participant. The task measuring auditory acuity (perceptual identification task) was administered at pre-treatment and post-treatment; the task measuring somatosensory acuity (oral stereognosis) was administered at pre-treatment except for those noted with an asterisk in the table.

Sensory tasks

Somatosensory acuity was measured with an oral stereognosis task in which children used their tongue tip to identify a letter embossed on a plastic strip. Due to the timing of task development and the unavailability of materials at both sites simultaneously, eleven participants were administered this task at pre-treatment, whereas 14 participants were administered this task at post-treatment. (See the discussion section for elaboration on this discrepancy.) Specifications for both stimulus materials and task administration were derived from Steele et al. (2014). The clinician handed the child the letter strip in the same orientation on each trial, and instructed the child to place the strip in their mouth just behind their top teeth. Children were told that the top of the letter would be toward the back of their mouth. They received ongoing instruction on this concept until they could correctly answer a comprehension question by indicating that the tip of a "V" would be toward the front of their mouths. After demonstrating comprehension of the orientation of the letters, children were provided a practice trial with no accuracy feedback. The children were instructed that the letters would all be capital letters and that the same letter could repeat. Letters ranging in size (2.5-8mm) were presented using an adaptive staircase paradigm, where size decreased following each correct responses and increased following each incorrect response. Following Steele et al. (2014), after 8 reversals in direction or after 28 trials (whichever occurred first), the score was calculated as the average letter size of only the correct responses; higher scores indicate a lower degree of

somatosensory acuity. *Table 4.1* shows mean letter size scores on this task for each participant.

To measure auditory acuity, we administered a forced choice identification task in which children labeled items from an acoustic continuum from *rake* to *wake* at both pre-treatment and post-treatment time points. Only pre-treatment auditory acuity was considered in the present analyses. A 9-step acoustic continuum from *rake* to *wake* was synthesized from a 10-year old TD female child's production of the word *rake*, as described in McAllister Byun and Tiede (2017). Children were administered all trials at a comfortable volume using over-the-ear headphones in a sound attenuated room. Following McAllister Byun and Tiede (2017), in each trial, the target item was presented two times with a 500-ms interstimulus interval. After presentation of each trial, children were instructed to indicate which word they heard by using a mouse to click on the corresponding word on a computer screen. Participants initially completed 8 practice trials in which the two endpoint stimuli were presented in a random order. The main task consisted of 8 randomly presented presentations of each of the 9 stimuli, for a total of 72 trials with a break at the halfway point. The proportion of *rake* responses was fitted to a sigmoidal logistic function, and the boundary width of the fitted function between the 25th and 75th percentile of probability of choosing *rake* was calculated. Wide boundary regions indicate reduced categorical labeling consistency, considered suggestive of poor auditory-perceptual acuity. There were three participants whose boundary widths were upper outliers (participant 16 at pre-treatment; participants 2 and 17 at post-treatment), indicating an inconsistent

response pattern. These values were coerced to 9.0 so that the size of the boundary region could not exceed the total size of the continuum. Previous researchers that have administered the same task to TD children have discarded upper outliers as potentially indicating reduced attention to the task (Cialdella et al., 2021; McAllister Byun & Tiede, 2017). However, children with RSE show larger average boundary regions than TD children (Cialdella et al., 2021), including a higher incidence of boundary widths exceeding the total size of the continuum. In the context of a study of children with RSE only, we elected to retain the upper outlier scores (n = 1 at the pre-treatment time point) as suggestive of a particularly high degree of difficulty with the auditory identification task. *Table 4.1* shows boundary widths at both pre-treatment and at post-treatment; the three outlier scores that were coerced to 9.0 are labeled with the original values in parentheses. Boundary widths did not differ from one another at the two time points (mean at pre-treatment = 3.01, mean at post-treatment = 2.35, $t(46.64) = 1.05$, $p = 0.297$).

Table 4.1. Child participant information and test scores, including evaluative tests and sensory probes.

Subject	Site	Gender	Age	PPVT-4 SS	GFTA-2 SS	MPT apraxia index	CTOPP-2 composite score	CELF-4 recalling sentences SS	Identification (boundary width)		Stereognosis (mean letter size)
									Pre	Post	
1	Haskins	F	11;9	110	110	0	110	12	2.7	1.8	6.2 *(post)*
2	Haskins	F	9;3	127	127	0	35	7	3.2	9.0 *(53.2)*	5.5 *(post)*
3	Haskins	F	9;0	114	70	1	90	10	4.5	2.6	7 *(post)*
4	Haskins	F	9;0	110	83	0	96	8	2.0	5.3	8 *(post)*
5	Haskins	F	10;10	103	77	0	92	10	2.6	2.0	7.3 *(post)*
6	Haskins	F	9;1	106	76	0	-	8	5.7	2.6	7.6 *(post)*
7	Haskins	M	9;6	118	82	1	112	16	5.8	2.8	5.6 *(post)*
8	Haskins	M	10;1	103	78	-	100	8	1.8	2.0	4.6 *(post)*
9	NYU	M	10;0	143	78	0	125	18	0.0	0.0	3.6 *(post)*
10	NYU	F	9;10	152	79	0	127	18	2.1	1.5	7.1 *(post)*
11	NYU	M	12;1	110	40	0	86	10	1.8	0.8	5 *(post)*
12	NYU	M	11;4	97	75	0	88	14	1.0	1.3	4.3 *(post)*
13	NYU	F	10;1	121	73	1	92	14	3.9	0.8	3.4
14	NYU	M	11;0	119	86	0	100	15	2.0	1.4	5.8 *(post)*
15	NYU	M	13;10	118	<40	0	94	14	1.3	0.0	6 *(post)*
16	NYU	M	9;1	113	76	1	100	12	9.0 *(47.2)*	4.6	6
17	NYU	M	14;7	101	69	0	65	8	3.7	9.0 *(13.7)*	4
18	NYU	M	10;10	113	74	0	94	15	1.7	0.6	7.8
19	NYU	F	11;1	132	78	1	110	11	1.3	0.0	7.5
20	NYU	M	9;3	137	83		84	13	2.8	3.3	5.5
21	NYU	F	10;1	139	68	0	100	17	1.9	1.1	5.1
22	NYU	M	10;4	129	73	0	105	10	2.1	0.0	6
23	NYU	M	12;7	100	71	0	96	16	0.0	2.1	5.4
24	NYU	M	10;7	112	77	0	107	10	1.4	2.5	7
25	NYU	F	9;6	114	67	0	88	16	5.9	1.6	5.7

Stimulability probe

A probe assessing stimulability in /ɹ/ production was administered at both pre-treatment and at post-treatment. Adapted from Miccio (2002), the stimulability probe involved direct imitation of 15 different syllables three times each in a standard order, for a total of 45 productions. The syllables included syllabic variants (/mɝ, dɝ, ɝg/), prevocalic variants in front vowel (/ɹi, ɹe, ɹaɪ/) and back vowel contexts (/ɹu, ɹo, ɹɑ/), and postvocalic variants in front vowel (/ɪɚ, ɛɚ, aɪɚ/) and back vowel contexts (/ɔɚ, ɑɚ, ɑʊɚ/). These syllables were elicited in the order listed here.

Audio recordings were collected in a quiet room with a mouth-to-microphone distance of five inches using a Zoom H4n recorder at NYU and a Sony PCM-M10/B 4 GB Voice recorder at Haskins. Recordings were digitized using a sampling frequency of 48,000 Hz and 16-bit encoding. Ultrasound images of the midsagittal tongue contour were co-collected with frontal video to identify productions in which the ultrasound probe was misaligned, as described in a later section. While the stimulability task elicited 45 productions, some productions were elicited multiple times to ensure that a clear ultrasound image was available; there were also exclusions made due to ultrasound probe misalignment.

Due to the relatively complex experimental setup, equipment failure and/or experimenter error led to some data loss. One participant (not listed in tables) was excluded entirely because frontal video, which was used to verify that ultrasound video was appropriately aligned, was not successfully obtained at either pre-treatment or post-treatment. In addition, there was no ultrasound data at pre-treatment for subjects 2 and 8, and there was no frontal video at pre-treatment for subjects 9, 10, 11, and 12. In these cases, the available data from the other time point was included because the data were missing at random (i.e., the probability of such exclusion was the same for all participants), allowing for the use of an unbalanced data set (Ibrahim & Molenberghs, 2009). Following the exclusion of one participant at both time points and the six exclusions at pre-treatment, there were a total of 25 participants with complete data for 19 pre-treatment time points and 25 post-treatment time points. This amounted to a total of 1950 productions across the 44 files.

Formant measurement

The first author measured the formant frequencies of the rhotic interval within each production using Praat software (Boersma & Weenink, 2019). Rhotic intervals were initially marked by trained students, as described in the section on ultrasound analysis. For each sound file, the first author determined optimal formant settings by visual inspection for each file for each participant (e.g., 5 formants in 5500 Hz). Using those formant settings, the author manually selected a point in the steady state portion of the rhotic interval judged to represent the minimum non-outlier value of the third formant, F3. A Praat script (Lennes, 2003) was used to extract the average first, second, and third formant frequencies in Hertz within a 14 ms window around the selected point.

The difference between the second and third formants in Hertz (F3-F2) was calculated from the raw acoustic measures for each token. To account for expected differences in formant frequencies related to age and gender, the F3-F2 values were normalized relative to age-matched peers using means and standard deviations from a published study of TD children (Flipsen, Shriberg, Weismer, Karlsson, & McSweeny, 2001; S. Lee, Potamianos, & Narayanan, 1999). Therefore, normalized F3-F2 distance was used as the acoustic index of rhoticity in the present study; previous research has indicated that this is the acoustic measure that best correlates with degree of perceived accuracy from expert raters (H. Campbell, Harel, Hitchcock, & McAllister Byun, 2018).

Reliability for the formant measurements was determined by a phonetically trained graduate student who followed the same procedure for

formant measurement described above for 10/44 files (23%). Reliability of the normalized F3-F2 distances between the original and remeasured files was assessed using intraclass correlation with single random raters. The calculated intraclass correlation of 0.96 indicates strong agreement between formant measurements performed by different individuals (Koo & Li, 2016).

Ultrasound data collection and processing

We recorded ultrasound video streamed from a Siemens Acuson X300 to a computer-based AverMedia video capture card. At NYU, a Siemens C8-5 wideband curved array transducer was used (frequency range 3.1–8.8 MHz, 25.6 mm footprint, 109 degree field of view) and at Haskins Laboratories, a C6-2 wideband curved array transducer (frequency range 1.8–6 MHz, 73.0 mm footprint, 90 degree field of view) was used. Scanning settings differed across participants and sites. At NYU, settings ranged across individuals from 30 to 45 frames per second with 7-11 cm zoom depth, while all participants at Haskins were imaged at 36 frames per second with 8 cm zoom depth. Adjustments in zoom depth were made by the clinician at NYU (the first author) to accommodate for children with different size vocal tracts so that the tongue would fill the ultrasound display.[6] The ultrasound video recordings were recorded at 60 frames

[6] Although it would be ideal to have identical settings across individuals, we do not regard the impact of these differences to be large enough to affect the present analyses for two main reasons. First, at even the lowest ultrasound frame rate of 30 frames per second, the selected ultrasound frame could be at most 33 ms from the true frame of interest, which is considered acceptable for the current non-dynamic analysis. Second, the curvature-based indices of ultrasound data

per second through the video capture card. The ultrasound probe was not stabilized relative to the child's head, but the transducer was placed in a microphone stand whose flex maintained reliable contact between the probe and the skin under the jaw, and the clinician supported alignment of the transducer relative to the head while monitoring image quality via the ultrasound display. Moreover, a post-processing procedure using frontal video was used to quantify the adequacy of probe alignment in each frame, as described toward the end of this section.

Ultrasound processing was completed by university students with training in phonetics and/or general linguistics. In the first phase of ultrasound processing, the students viewed the waveforms and spectrograms of each sound file in Praat (Boersma & Weenink, 2019) in order to mark each rhotic interval in a time-synced TextGrid file. The /ɹ/ boundaries were placed at the onset and the offset of any formant transitions, thereby including the entire region characterized by a lowered F3, and consequently, a reduced F3-F2 distance.

The second phase involved tracing the tongue shape in the ultrasound image within the intervals determined by Praat TextGrids using the GetContours program (Tiede, 2020). Once the program indicated that they were viewing frames within a target interval, the students stepped incrementally through the interval to select the ultrasound frame that most clearly represented the maximal

---

(described below) can still be measured without knowledge of spatial orientation (Ménard et al., 2012; Stone, 2005), which includes size differences associated with depth variation.

tongue constriction for an /ɹ/ target. They were encouraged to select frames near the temporal midpoint of the interval but were informed that non-midpoint frames could be selected as needed to avoid regions of poor image quality. Students were also familiarized with the variety of tongue shapes that can be used to achieve typical /ɹ/, and were made aware that tongue shapes from misarticulated productions would be expected to deviate substantially from those expected shapes.

After selecting the target frame, the students traced or 'tagged' the contour by placing sixteen spline anchor points just under the visible bright line representing the tongue surface. The manually-placed points were then redistributed across the spline into 100 evenly-spaced contour points. After this initial tagging, a preliminary round of quality assurance was carried out in which another student with specialized training in phonetic analysis checked all files for consistency and remeasured any tokens that did not meet the standards described below. *Figure 4.1* shows sample sets of sixteen anchor points from the same participant for an incorrectly articulated /ɹ/ target at pre-treatment and a correctly articulated /ɹ/ target at post-treatment. The top image shows incorrect /ɹ/ in "ra" at pre-treatment; the bottom image shows correct /ɹ/ in "ra" at post-treatment. Both images are from subject 21. The first author extracted the contour coordinates from each traced target frame. From these 100 points, custom scripts were used to calculate MCI (Dawson, 2016) and NINFL (ComputeCurvature) from the contour coordinates.

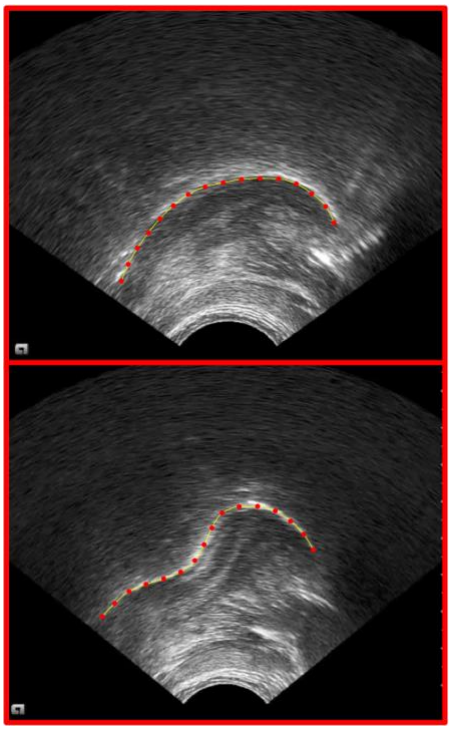

Figure 4.1. Sample sets of sixteen evenly distributed anchor points traced in GetContours.

As the primary means of quality assurance, the frontal video was used to identify tokens produced with an excessive amount of angular or lateral displacement from the desired midsagittal alignment of the transducer. To do this, adhesive blue dots placed along the vertical midline of the transducer and the child's face were tracked in Matlab (MathWorks Inc., 2000) through successive frames using an automated procedure, resulting in a set of coordinates relating the position and orientation of the probe to the child's head. The common acoustic signal was then used to align frames from the ultrasound video and the frontal video, such that for each aligned, angular displacement (in degrees) and lateral displacement (in millimeters) of the facial dots relative to the transducer dots could be calculated. Frames with more than 1.5 standard deviations of lateral

displacement (15.8 mm) or 1.5 standard deviations of angular displacement (12.4 degrees) were discarded as inaccurately sampling the midsagittal tongue. Across all 25 stimulability files processed for the present study, 5.9% (116 /1950) of frames were flagged due to angular displacement and 3.8% (75/1950) of frames were flagged due to lateral displacement. This resulted in a total of 146 tokens discarded, as 45 of the tokens were flagged for both angular and lateral displacement; a total of 1804 tokens remained after removal of misaligned frames. For more detail on this procedure, see Kabakoff et al. (2021).

As a final quality check, the first author scaled the extracted coordinates within each participant and visualized the scaled coordinates in R (R Core Team, 2019) using RStudio (RStudio Team, 2019). Through this process, all tokens were judged to represent complete contours. Any token with a NINFL value exceeding five (n = 10) was removed according to the criterion introduced in Preston et al. (2019) and followed in Kabakoff et al. (2021). Similarly, any token with an MCI value exceeding six (n = 16) was removed based on the distribution of values presented in Dawson et al. (2016), leaving a total of 1778 productions in the final data set. *Table 4.2* reflects these final counts of tokens by subject at each time point after all data cleaning was completed.

Subject	Pre	Post
1	44	35
2	-	20
3	45	44
4	45	44
5	45	24
6	37	43
7	45	44
8	-	44
9	-	40
10	-	41
11	-	45
12	-	44
13	43	45
14	43	33
15	38	38
16	38	36
17	25	38
18	42	37
19	41	40
20	39	52
21	43	44
22	45	40
23	33	42
24	39	45
25	45	45
Total:	**775**	**1003**

Table 4.2. Final number of tokens for each participant on the stimulability probe.

Reliability was tested on 16% of the files (n = 7) in the cleaned data set for both MCI and NINFL. Following the precedent set in Preston et al. (2019), all tongue contours for each file were retagged on the previously identified target frame by a second trained tagger. For the continuous MCI measure, the intraclass coefficient (ICC) with single random raters was used, and ratings were compared with the standards published for this statistic in Koo and Li (2016). For the ordinal NINFL measure, Cohen's kappa was used and ratings were compared

with the standards published for this statistic in McHugh (2012). The intraclass coefficient (ICC) for MCI was 0.71, indicating moderate agreement, and Cohen's kappa for NINFL was 0.33, indicating fair agreement.[7]

Analyses

Our first question was whether tongue complexity is associated with acoustically measured production accuracy for /ɹ/ targets. To address this question, we fit linear mixed-effects regression models predicting acoustic production accuracy (normalized F3-F2 distance) from a co-regressor of tongue complexity (either MCI or NINFL) and a fixed effect of treatment time point (pre/post). The interaction between tongue complexity and treatment time point was also included, justified on the theoretically motivated hypothesis that the relationship between between tongue complexity and production accuracy could differ before versus after treatment. We also included random intercepts for child and syllable type. Sensory measures (auditory and somatosensory acuity) were

---

[7] Reliability was initially calculated as poor for MCI (ICC = 0.31) and fair for NINFL (kappa = 0.32). Because these reliability values were low relative to previous results following the same procedure from Kabakoff et al. (2021) following the same procedure, the original and reliability files were inspected by the first author, who identified a discrepancy in how points at the very edge of the tongue tip were handled. This led to the realization that the students performing reliability assignments had been given slightly different instructions on how to handle the anterior edges of the tongue contours than the original taggers. The students were re-trained with instructions consistent with those provided to the original taggers and reviewed the complete set of reliability files, making adjustments as needed to adhere to the modified convention. The reliability results reported are after these adjustments. The low initial reliability and need for re-training represents a limitation of the present study, which will be revisited in the discussion.

not included in these models because of the potential for multicollinearity, as the sensory measures were hypothesized to correlate with tongue complexity.

The second question was whether a child's degree of somatosensory acuity predicts tongue complexity for /ɹ/ targets. We addressed this question by fitting a linear mixed-effects regression model (for MCI) and an ordinal mixed-effects regression model (for NINFL) predicting tongue complexity from somatosensory acuity and including a fixed effect of treatment time point (pre/post). In these models predicting tongue complexity, we controlled for auditory acuity and included the interactions between somatosensory acuity and treatment time point and between auditory acuity and treatment time point on the theoretically motivated basis that the relationships between sensory acuity and tongue complexity could differ before versus after treatment. We also included random intercepts for child and syllable type. We ran these as separate models with either MCI or NINFL as the outcome variable.

Mixed effects models were fitted following recommendations presented in Harel and McAllister (2019). All linear mixed effects models were performed using the 'lme4' package (Bates et al., 2015), and ordinal mixed-effects regression were performed using the 'clmm' package (Christensen, 2015). Complete data and code to reproduce all figures and analyses in the paper can be retrieved at https://osf.io/78zqb/.

Results

1) Does tongue complexity predict perceptual rating of accuracy for /ɹ/ targets?

The models used to address this question involved predicting normalized F3-F2 distance (acoustically measured accuracy, where a smaller normalized F3-F2 distance indicates higher accuracy) from tongue complexity (MCI or NINFL), treatment time point (pre/post), their interaction, and random intercepts for child and syllable type. Full model outputs are presented in *Table 4.3*, with the model predicting accuracy from MCI (centered around the mean) on the left-hand side, and the model with NINFL on the right. The pre-treatment time point served as the reference level for both models. The top section shows that at pre-treatment, there was a significant positive association between MCI and normalized F3-F2 distance, but not between NINFL and normalized F3-F2 distance. The middle section shows change over time, with the negative coefficients indicating that the mean normalized F3-F2 distance was significantly lower at post-treatment than at pre-treatment in the models for both MCI and NINFL. The bottom section of the table shows that in the model with MCI as a predictor, there was a significant

	Linear Mixed Effects Model with MCI as a Predictor	Linear Mixed Effects Model with NINFL as a Predictor
Intercept (pre at mean MCI/NINFL)	**β = 8.17, SE = 0.61, p < 0.0001***	**β = 7.98, SE = 0.64, p < 0.0001***
Tongue complexity (slope for pre)	**β = 0.58, SE = 0.15, p = 0.00016***	β = -0.030, SE = 0.087, p = n.s.
Treatment time point (post vs pre)	**β = -6.23, SE = 0.13, p < 0.0001***	**β = -6.27, SE = 0.38, p < 0.0001***
Tongue complexity × treatment time point (slope for post vs pre)	**β = -1.01, SE = 0.18, p < 0.0001***	β = 0.064, SE = 0.14, p = n.s.
Random intercept subject	Variance = 9.012, SD = 3.00	Variance = 9.36, SD = 3.059
Random intercept syllable type	Variance = 0.059, SD = 0.243	Variance = 0.045, SD = 0.21

Table 4.3. Output for model predicting normalized F3-F2 distance from MCI and NINFL, treatment time point (pre/post), and the interaction between these predictors.

interaction between tongue complexity and treatment time point, indicating that the slope relating MCI to F3-F2 distance differed at post-treatment relative to pre-treatment. Specifically, while higher MCI values were associated with higher normalized F3-F2 distances at pre-treatment, they were associated with lower normalized F3-F2 distances at post-treatment. This interaction was not significant in the model with NINFL as a predictor. *Figure 4.2* shows normalized F3-F2 distance on the y-axis and tongue complexity on the x-axis. Both panels show a difference in acoustically measured accuracy at pre- vs post-treatment time points, where normalized F3-F2 distances are smaller in accurate targets. The left panel also shows the significant positive association between MCI and normalized F3-F2 distance at pre-treatment, and the significantly more negative association between these predictors at post-treatment. The lack of any association between acoustically measured accuracy and NINFL can be visualized in *Figure 4.2* (right panel). To avoid overlapping values, jitter is added to the ordinal NINFL plot. Smoothed regression lines depict 95% confidence intervals.

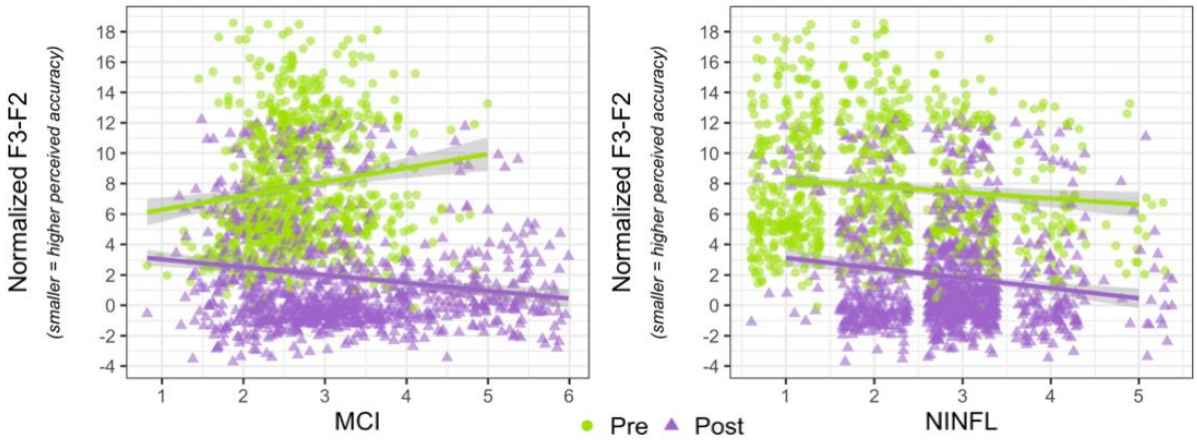

Figure 4.2. Acoustically measured accuracy versus tongue complexity, separated by treatment time point (pre/post).

2) Does degree of somatosensory acuity predict tongue complexity for /ɹ/ targets?

The models used to address this question involved predicting tongue complexity from somatosensory acuity (centered around the mean), treatment time point (pre/post), and the interaction between somatosensory acuity and treatment time point; auditory acuity (centered around the mean) was included as a covariate, as well as the interaction between auditory acuity and treatment time point. The model included random intercepts for child and syllable type. *Table 4.4* presents results from the model predicting MCI (left side) and for the model predicting NINFL (right side). The top section shows that tongue complexity was significantly higher at post-treatment than pre-treatment based on both tongue complexity metrics. This effect can be visualized in *Figure 4.3*, which depicts MCI in the left panel and NINFL in the right panel. The middle section of the

	Linear Mixed Effects Model Predicting MCI	Ordinal Mixed Effects Model Predicting NINFL	
Intercept (pre at mean somatosensory/auditory acuity)	**β = 2.75, SE = 0.14, p < 0.0001***		
Treatment time point (post vs pre)	**β = 0.44, SE = 0.035, p < 0.0001***	**β = 1.77, SE = 0.11, p < 0.0001***	
Somatosensory acuity (slope for pre)	β = -0.068, SE = 0.11, p = n.s.	β = 0.098, SE = 0.11, p = n.s.	
Auditory acuity (slope for pre)	β = 0.037, SE = 0.068, p = n.s.	β = -0.024, SE = 0.068, p = n.s.	
Somatosensory acuity × Treatment time point (slope for post vs pre)	**β = -0.17, SE = 0.029, p < 0.0001***	β = -0.011, SE = 0.081, p = n.s.	
Auditory acuity × Treatment time point (slope for post vs pre)	**β = -0.11, SE = 0.017, p < 0.0001***	β = -0.054, SE = 0.046, p = n.s.	
Random intercept subject	Variance = 0.043, SD = 0.66	Variance = 0.33, SD = 0.57	
Random intercept word	Variance = 0.0027, SD = 0.052	Variance = 0.020, SD = 0.14	
Threshold coefficients:		1|2	β = -1.86, SE = 0.15
		2|3	β = 0.90, SE = 0.15
		3|4	β = 3.24, SE = 0.17
		4|5	β = 5.24, SE = 0.22

Table 4.4. Output for model predicting MCI and NINFL from treatment time point (pre/post), somatosensory acuity, auditory acuity, and the interactions between treatment time point and somatosensory acuity and between treatment time point and auditory acuity.

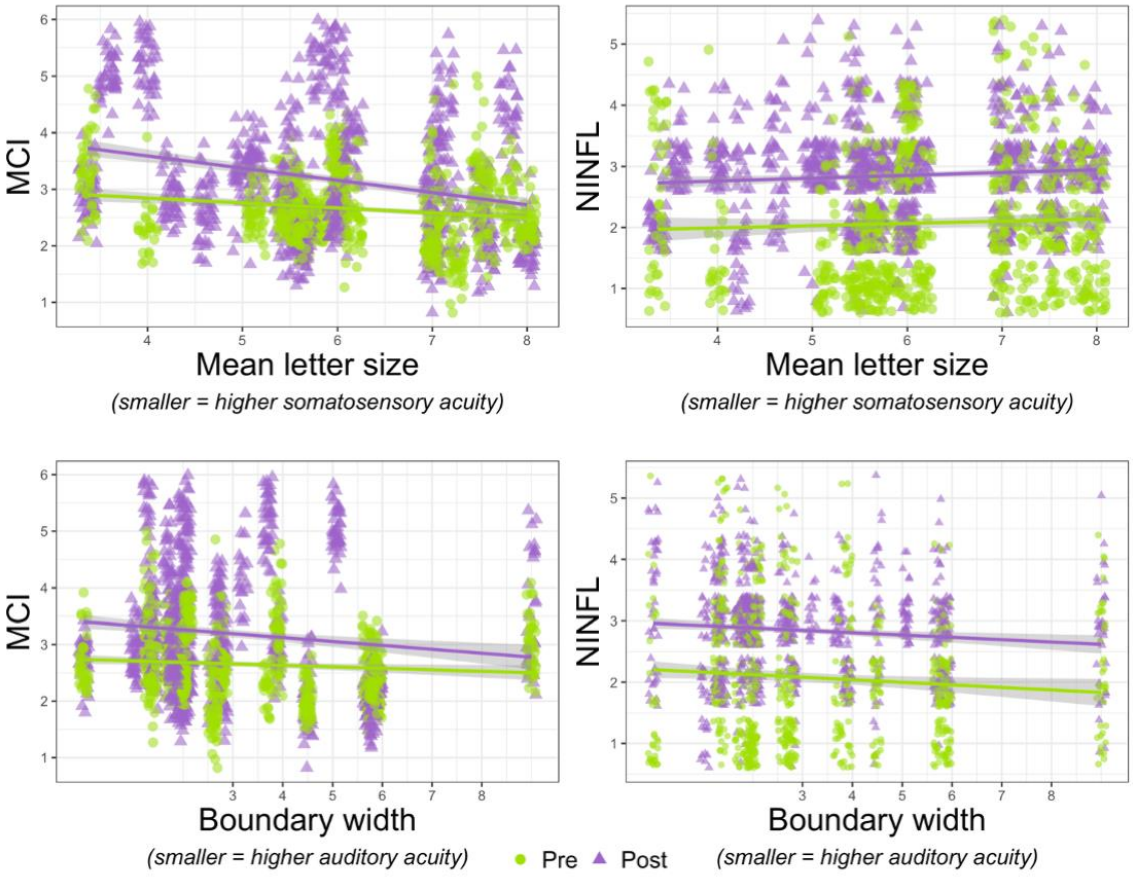

Figure 4.3. Tongue complexity by somatosensory acuity and treatment time point (pre/post) and by auditory acuity and treatment time point (pre/post).

table shows that, for the pre-treatment reference time point, there was no association between somatosensory acuity and tongue complexity based on either measures. Likewise, at the pre-treatment time point, there was no association between auditory acuity and tongue complexity. The bottom section shows significant interactions between somatosensory acuity and time, as well as auditory acuity and time, in the model predicting MCI only. For both somatosensory and auditory measures, the association between sensory acuity and tongue complexity was significantly more negative at post-treatment than at the pre-treatment time point. Since smaller scores on the sensory tasks are indicative

of stronger acuity, these results suggest that stronger sensory ability was associated with higher tongue complexity at the post-treatment time point only. These interactions were not statistically significant, compared to a cutoff of 0.05, when predicting NINFL. *Figure 4.3* depicts tongue complexity by somatosensory acuity and time point (pre/post) in the top panels and by auditory acuity and time point (pre/post) in the bottom panels. Plots for MCI are shown in the left panels and plots for NINFL are shown in the right panels. To avoid overlapping values, jitter is added to the ordinal NINFL plots. Somatosensory acuity is plotted as mean letter size on the stereognosis task and auditory acuity is plotted as boundary width on the perceptual identification task, such that each measure is inversely related with degree of sensory acuity. Smoothed regression lines depict 95% confidence intervals. The significant interactions based on MCI can also be visualized in the figure, including the negative association between mean letter size and tongue complexity at post-treatment (top left panel) and the negative association between auditory acuity and tongue complexity at post-treatment (bottom left panel).

Discussion

The present study explored relationships among tongue complexity, somatosensory acuity, and acoustically measured production accuracy of rhotic targets in children with RSE on a stimulability task administered both before and after treatment targeting rhotic misarticulation. As expected, children showed higher acoustically measured production accuracy (smaller normalized F3-F2 distance) and higher tongue complexity at post-treatment than at pre-treatment.

Our first question asked whether higher tongue complexity, which we use as an index of motor skill, would be associated with more accurate production of rhotic targets. In the model for MCI only, there was a significant interaction such that higher tongue complexity was associated with higher normalized F3-F2 distance at pre-treatment (lower accuracy), but this association was reversed at the post-treatment time point, as shown in *Figure 4.2*. The second question asked whether tongue complexity for rhotic targets would be higher for children with stronger somatosensory acuity (i.e., smaller mean letter size on the oral stereognosis task), while controlling for auditory acuity. Based on MCI only, there were significant interactions between somatosensory acuity and time point, as well as between auditory acuity and time point. In both cases, stronger sensory performance was associated with higher tongue complexity at the post-treatment but not at the pre-treatment time point.

Implications

Although the effects of treatment were not the focus of the present study, we first briefly comment on the robust differences between the pre-treatment time point and the post-treatment time point across models. That is, both acoustically measured production accuracy (as indicated by lower F3-F2 distance) and tongue complexity (based on both measures) were found to be higher after treatment than before treatment. This serves as the first replication of the result from Preston et al. (2019) that school-aged children with RSE producing /ɹ/ had higher NINFL values at post-treatment relative to pre-treatment. That study also observed higher NINFL values in productions that were rated perceptually correct relative to

incorrect productions, a finding that was indirectly replicated in the present study via acoustically measured accuracy.

The observation that acoustically measured accuracy on the stimulability probe was significantly higher at post-treatment than at pre-treatment provides a suggestion that the children acquired the changes in tongue shape from their experience in the ultrasound-based treatment program. Because there was no control condition in the present study to enable us to rule out the possibility of spontaneous improvement due to maturation, however, it is not possible to draw any strong conclusion regarding the effect of treatment. Nonetheless, spontaneous improvement past the age of eight is considered unlikely for children with RSE (Gibbon & Paterson, 2006). Therefore, the current results suggest that treatment was a likely contributor to the higher tongue complexity and the higher accuracy observed at the post-treatment time point.

In response to our first question, MCI showed a significant negative association between tongue complexity and acoustically measured production accuracy (i.e., a positive association with normalized F3-F2 distance) at pre-treatment; however, the same measure indicated that this relationship was significant and positive at post-treatment. Based on past research showing that perceptually accurate productions of /ɹ/ are associated with higher tongue complexity than less accurate productions (Kabakoff et al., 2021; Preston et al., 2019), we had predicted that tongue complexity would be positively associated with acoustically measured production accuracy. The reverse relationship was observed before treatment, indicating that at this time point, rhotic targets

produced with more complex tongue shapes were associated with lower degrees of production accuracy. However, the expected relationship was observed at post-treatment, suggesting that after treatment, complex tongue shapes were associated with greater degrees of production accuracy. We interpret this treatment-based difference in the relationship between tongue complexity and production accuracy with caution because of the possibility of chance fluctuations in our relatively small sample of children with RSE. If the current finding is found to be robust in future research, it could reflect a pattern in which some children with RSE have developed complex tongue shapes for rhotic targets as a maladaptive response to prior treatment. That is, in response to verbal and visual descriptions of typical /ɹ/ articulation provided during conventional /ɹ/ treatment, they may have been trained to produce something resembling a complex tongue shape characteristic of /ɹ/, but they have not succeeded in fine-tuning their production pattern to achieve the desired acoustic and perceptual consequences. Then, over the course of ultrasound intervention, they learned to adjust their tongue shapes to produce acoustically and perceptually accurate /ɹ/. This is an admittedly speculative interpretation, particularly in the context of the current group of participants with heterogeneous treatment histories. However, there have been previous reports of children with RSE who produced /ɹ/ with a complex tongue shape but inaccurate perceptual quality, and then learned to refine their tongue positioning to produce perceptually accurate /ɹ/ through ultrasound-based treatment (Boyce, 2015).

Our second question sought to determine whether performance on an oral stereognosis task (as an index of somatosensory acuity) predicts degree of tongue

complexity in children with RSE. We had hypothesized that children with stronger somatosensory acuity would have more complex tongue shapes for the target phoneme /ɹ/, consistent with research demonstrating a relationship between somatosensory acuity and speech outcomes (Fucci & Robertson, 1971; Kabakoff et al., 2020; McNutt, 1977). This predicted relationship between somatosensory acuity and tongue complexity was not observed at the pre-treatment time point. However, the significant interaction between time point and somatosensory acuity suggested that suggested that higher responsiveness to treatment, as indexed by greater observed tongue complexity, was predicted by stronger stereognosis scores. Also recall that we had included auditory acuity as a covariate that we controlled for in the model because we wanted to distinguish its relationship with tongue complexity from the relationship between somatosensory acuity and tongue complexity. There was a significant interaction between time point and auditory acuity suggesting that after treatment only, a smaller boundary width on the perceptual identification task (i.e., stronger auditory acuity) was associated with a greater degree of tongue complexity, as predicted. This result replicates the observation from Cialdella and colleagues' (2021) finding of no association between pre-treatment auditory acuity and accuracy. However, Preston et al. (2020) and Cialdella et al. (2021) did find an association between auditory acuity and magnitude of treatment response, although for females only in the latter study. Our finding that greater auditory acuity was associated with higher degrees of tongue complexity after treatment is broadly compatible with these previous results, particularly in light of our current finding that productions with greater

tongue complexity after treatment were also judged to be more accurate based on acoustic measures. Although we interpret these treatment-based differences in the relationships between sensory acuity and tongue complexity with caution, the interactions could reflect a phenomenon in which children with RSE who have strong sensory acuity in at least one domain are more able to achieve high complexity tongue shapes through ultrasound-based biofeedback treatment.

Comparison of MCI versus NINFL

For both of the current questions, our findings based on MCI were more compatible with our hypotheses than those based on NINFL. This was surprising given that both MCI and NINFL are intended to measure tongue shape complexity and that the two measures were moderately correlated with one another across the data set ($r(1776) = 0.43$, $p < 0.00001$). Furthermore, both measures were found to be higher after treatment than before treatment, as predicted, which suggests that both have the potential to be valid representations of tongue complexity. For our first question, MCI showed the expected relationship between tongue complexity and acoustically measured accuracy at post-treatment, although it showed the reverse relationship at pre-treatment. However, NINFL did not indicate any relationship between tongue complexity and acoustically measured accuracy. There was not even a trend in the predicted direction; instead, the non-significant coefficients from the model with NINFL as a predictor suggested that the pre- and post-treatment relationships were reversed for NINFL relative to MCI. For our second question, MCI showed the expected relationship between somatosensory acuity and tongue complexity, as well as

between auditory acuity and tongue complexity, at post-treatment. However, NINFL did not show any relationships between sensory acuity and tongue complexity.

The present findings are particularly surprising given that in a previous study of younger children with and without SSD (Kabakoff et al., 2021), predicted differences in complexity for /ɹ/ were observed in the analyses using NINFL but not MCI. That is, based on NINFL only, tongue complexity for /ɹ/ was found to be higher in TD children than in children with SSD, and also higher in correct versus incorrect productions, independent of diagnosis. As discussed in Kabakoff et al. (2021), the computational differences between the metrics may have led to these differential patterns with MCI versus NINFL. That is, it is possible to have a high MCI value with only one point of inflection if the local curvature is high (i.e., the radius of curvature is small) at some point, such as near the tongue tip. This could result in high MCI values but low NINFL values for retroflex tongue shapes, which often involve tight local curvature at the raised tongue tip. In contrast, a bunched shape would not have high local curvature, and therefore may have relatively lower MCI values but higher NINFL values.[8] In our qualitative examination of both data sets, it was noted that many of the children in the present study favored retroflex tongue shapes for /ɹ/, possibly as a result of the tongue shaping cues that were provided in treatment. (While both bunched and

[8] However, this hypothesis is not consistent with recent work from Heyne, Wang, Derrick, Dorreen, and Watson (2020), which found that higher MCI values were more likely to be associated with tip-down than with tip-up rhotic variants in adults.

retroflex shapes were offered as models in treatment, a relatively large portion of children were observed to respond positively to cues for retroflex shapes, leading to subsequent sessions reinforcing that tongue shape.) The younger children in Kabakoff et al. (2021) did not receive ultrasound biofeedback treatment, and they were subjectively judged to produce more bunched shapes than the older children, consistent with estimates of the general prevalence of these tongue shapes in American English speakers (Mielke, Baker, & Archangeli, 2016). Follow-up research efforts are underway to quantify the proportion of retroflex and bunched tongue shapes in each sample and examine the relationship between MCI, NINFL, and rhotic tongue shape category.

Next steps and limitations

We used MCI and NINFL as tongue complexity metrics in the current study because previous research indicated that these metrics are useful for distinguishing tongue shapes in children (Kabakoff et al., 2021; Preston et al., 2019). However, we acknowledge two limitations of these measures. First, as discussed in Kabakoff et al. (2021), both metrics represent tongue shape complexity from only one midsagittal cross-section, thus ignoring lingual interactions with hard structure morphology (Brunner et al., 2009) and potentially relevant parasagittal complexity associated with lateral bracing (Gick et al., 2017). To address this limitation, future research should consider multiple cross-sections simultaneously with reference to hard structures, as is possible with three-dimensional ultrasound (Lulich & Pearson, 2019). Second, both MCI and NINFL may have suboptimal reliability when different individuals trace

ultrasound images. As described, we attributed the reduced reliability that we initially observed for MCI to be associated with a specific difference in the instructions provided for handling tracing of the tongue tip. The decision of how far the tongue contour is traced both posteriorly and anteriorly can significantly affect the calculated degree of tongue complexity. To avoid such differences, it is important to provide training that is both thorough and standardized so all individuals follow the same conventions for tracing contours. We also acknowledge that reliability remained relatively low (Cohen's kappa = 0.33, considered fair agreement) for NINFL even after students were retrained to follow the same conventions. This reduced reliability may be related to the fact that computation of NINFL requires setting of both a filter cutoff and threshold that together serve to determine which changes in curvature count as meaningful inflections, as described in Kabakoff et al. (2021). Although the current settings were the same ones used in Kabakoff et al. (2021), it is possible that optimal settings for this metric may differ between older versus younger children; future research should consider optimizing the settings for each population. Additionally, idiosyncratic differences in manual tagging may lead to differences in the number of inflections represented in this ordinal metric. To avoid the complications introduced by manual tagging, future research should consider the use of automated tongue contour tracing (Laporte & Ménard, 2018; Li, Kambhamettu, & Stone, 2005).

An additional limitation of the present pilot study pertains to the task used to measure somatosensory acuity. An oral stereognosis task was used because of

its relative ease of administration and because previous research had shown reduced performance on this task for children with SSD relative to TD peers. However, it may be that a task tapping proprioceptive aspects of somatosensory acuity would be more appropriate for the present investigation. Proprioception may be particularly important to consider in the context of rhotic targets, which feature a complex tongue shape but involve relatively limited linguopalatal contact. Likewise, a perceptual identification task was used to measure categorical labeling consistency as an index of auditory acuity. However, previous studies showing a correlation between auditory acuity and distinctness of speech contrasts have used thresholds from auditory discrimination tasks (Franken, Acheson, McQueen, Eisner, & Hagoort, 2017; Ghosh et al., 2010; Perkell et al., 2004). Given this literature precedent, it is possible that a discrimination threshold may be better than a categorical labeling consistency score from an auditory identification task at representing the ability to use auditory feedback to update motor plans. Future research should consider tasks probing multiple dimensions of sensory skill in order to provide an optimal representation of sensorimotor profiles as they relate to speech production.

A related limitation of this pilot study is that some of the participants' somatosensory acuity scores were collected at pre-treatment, while others were collected at the post-treatment evaluation or briefly after the end of treatment. The reason for this discrepancy was that the oral stereognosis task was still in development at the start of the treatment study from which we drew our measurements. Furthermore, a single set of materials for administration of the

task was shared across the two sites participating in the study. As a result, a small number of participants at the NYU site, as well as the majority of participants at the Haskins site, completed the stereognosis task only after the end of treatment. While we acknowledge that these differences in the timing of task administration were not ideal, we did not hypothesize that children would improve on a somatosensory task as a result of treatment. Still, future research should better control the timing of administration of the sensory tasks.

A final consideration pertains to the fact that we estimated degree of perceived accuracy using an acoustic measure in the context of a stimulability probe. It is worthwhile to reflect on the rationales for the decision to use a stimulability task and the decision to use acoustic ratings as the index of perceptual accuracy. A stimulability probe was selected for analysis in the present study because the visual and auditory models provided in that context provide a maximal level of support for the production of target sounds. However, improvement on a stimulability task may not reflect robust treatment gains that would generalize to productions in contexts in which less clinician support was provided. Although measuring treatment effects was not a primary focus of the present paper, we did attempt to draw inferences about treatment response from the association between time point (pre- versus post-treatment) and acoustically measured accuracy. Future research should investigate whether the potential treatment-based associations observed in the present study are robust, particularly by analyzing productions from a task that better reveals generalization of treatment gains, such as a word probe containing /ɹ/ in words that were not

targeted in treatment. In a similar vein, acoustic measurements were analyzed in order to provide a continuous measure of perceived accuracy, which therefore is equipped to reveal very small differences in rhotic production. However, such fine-grained measures may have reduced ecological validity when interpreted in connection with treatment effects. Taken together, although the decision to use acoustic ratings in a stimulability task equipped the current study to reveal even small changes in production, in future research it would be advantageous to also consider human listener ratings in the context of a word probe intended to measure generalization.

Clinical significance

The significant association between tongue complexity and acoustically measured accuracy at both pre-treatment and post-treatment points to the potential clinical importance of evaluating tongue shape patterns prior to treatment. The long-term clinical goal of this program of research is to be able to identify how information about a child's sensory and motor profile may inform who might respond to various forms of intervention. However, our results are suggestive of the possibility that some children with RSE may have developed complex tongue shapes that were not associated with production accuracy, but through targeted intervention, learned to adjust those shapes to align better with accurate percepts. If this interpretation is supported by future evidence, it might suggest that children with *high* tongue shape complexity at pre-treatment are likely to be good candidates for ultrasound-based treatment. This conclusion would be incongruent with our original prediction that children with *low* tongue shape complexity at

pre-treatment would benefit most from this kind of treatment. Alternatively, instead of one of these possibilities being true, it may be that either excessively high or excessively low tongue complexity values could be suggestive of issues with motor skill. Whichever of these scenarios is supported by further research, all are compatible with the idea that tongue complexity could be a useful index of motor involvement for children with RSE. However, as tongue complexity may not fully represent motor skill as it relates to production outcomes in this population, future research should also consider other known indices of motor skill, including articulatory coupling (Goffman & Smith, 1999; Green et al., 2000) and linguopalatal contact patterns (Fletcher, 1989; Gibbon, 1999), and coarticulation (Noiray et al., 2018).

When including both sensory measures in our statistical models, there was no significant association between either sensory acuity measure and tongue complexity at the pre-treatment time point. However, the finding of significant relationships between both sensory measures and tongue complexity at the post-treatment time point suggests that participants with stronger sensory acuity may have also had greater response to treatment. That is, children with a higher degree of sensory acuity may have been better able to learn associations between complex tongue shapes and perceptually accurate outcomes during targeted intervention. Furthermore, because post-treatment patterns suggested that both somatosensory and auditory acuity were associated with tongue complexity and that tongue complexity was associated with degree of acoustically measured accuracy, it follows that somatosensory and auditory acuity could be direct

predictors of degree of production accuracy at post-treatment. While this pilot study indirectly establishes this as a hypothesis, future research should directly explore the connection between somatosensory and auditory acuity and degree of perceived accuracy over the course of treatment; such associations could be valuable for clinical assessment and treatment planning.

Identifying children's specific areas of strength and weakness in the context of sensorimotor learning could help clinicians identify the treatment approach that is most likely to be beneficial. This goal is aligned with the concept of "personalized learning" (Perrachione, Lee, Ha, & Wong, 2011; Wong & Perrachione, 2007), which proposes that tailoring training to address an individual's specific deficit areas can help maximize learning outcomes. Applying a personalized learning framework to the treatment of RSE in the context of this research, children with reduced auditory acuity may be best-suited to a form of treatment that incorporates an auditory-perceptual training component. More relevant to the present study, our results suggest that children with high somatosensory acuity may respond positively to ultrasound biofeedback. Furthermore, although we had predicted that children with weak motor skill would benefit from ultrasound biofeedback treatment, our results suggested that children with high tongue complexity at baseline may also benefit from this form of treatment. With a future goal to optimize selection of personalized learning for children with RSE, we acknowledge that cognitive factors, phonological awareness, and personal preferences should also be considered, beyond the sensorimotor skills focused on in this research. Future research should therefore

consider the multidimensional mechanisms behind response to various treatment types in children with diverse sensorimotor profiles.

Acknowledgments

This research was supported by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under Grant F31DC018197 (H. Kabakoff, PI), Grant R01DC013668 (D.H. Whalen, PI), and Grant R01DC017476 (T. McAllister, PI). Additional support was provided through an Acoustical Society of American Stetson Scholarship and an American Speech-Language-Hearing Foundation New Century Scholars Doctoral Scholarship. We thank Siemens Medical Solutions USA, Inc., for making their Acuson ultrasound scanner available for this project. We gratefully acknowledge Emily Phillips for treatment administration at Haskins Laboratories, Sam Pearl Beames for processing video to measure ultrasound probe alignment, and Amanda Eads for completing formant measurement for the reliability analysis. We also extend gratitude to the ultrasound image tracing and reliability leadership team, including Graham Tomkins Feeny, Sam Pearl Beames, and Zhigong Ma.

CONCLUDING REMARKS

The overall goal of this dissertation was to quantify specific dimensions of sensorimotor skill in children and to determine whether these skills are associated with differences in speech outcomes. The first sections below address each of the aims in light of the findings from the three manuscripts. The subsequent sections include discussion of the primary applications of this work, including the contributions to basic science as well as the implications for assessment and treatment of clinical populations. Limitations and future directions for follow-up research will then be discussed.

Aim 1: Determine whether there are differences in tongue complexity in individuals known to differ in speech production abilities

The first manuscript addressed the aim of examining differences in tongue complexity between adults and young children through qualitative comparisons in productions of multiple phonemes. This summary focuses on the results for MCI and NINFL, the two measures that were judged to most directly reflect differences in tongue shape patterns across the two samples. For adults, tongue complexity measures were low for vowels and high for /ɹ/ and /l/, and tongue complexity values broadly ranked individual phonemes in a pattern that aligned with pre-established complexity classes, although there was a moderate amount of overlap of tongue complexity values across phonemes. For children, there was substantially more overlap in tongue complexity values across phonemes than for adults. Child tongue complexity values were generally low for the vowel /æ/ and high for /ɹ/, but patterns for /ɪ/ and /l/ did not align with expectations based on

adult patterns. Specifically, /ɹ/ was classified as having medium complexity based on both metrics while /l/ was classified as having low complexity based on MCI. Despite these unexpected findings, the presence of differences between earlier-developing vowels like /æ/ and late-developing approximants like /ɹ/ suggest that MCI and NINFL can be used to reveal differences in tongue complexity between phonemes and phoneme classes in both adults and children. Furthermore, they suggest that examination of child tongue shape complexity patterns may reveal differences from adult tongue shape patterns for various targets. Although the late-developing sounds /l/ and /ɹ/ were highlighted in this dissertation, differences between adults and children on other targets included the fact that alveolar stops were classified as having high complexity for adults based on one metric, whereas both metrics classified /t/ as having low complexity for children. For /w/, divergent values were observed in both age groups, such that one metric classified /w/ as having relatively high tongue complexity for both adults and children, but the other metric placed /w/ in the middle of the complexity range for adults and at the bottom of the range for children. Although the tongue complexity measures do not always agree with expectations across all targets, the most notable contribution from this study is that the current tongue complexity measures can reveal differences in tongue complexity between early-and late-developing phonemes and phoneme classes in children.

The second manuscript quantitatively addressed Aim 1 by determining whether there are differences in tongue complexity based on age, disorder status, and perceived accuracy in a sample of younger children ages 4-6. Although age

did not predict tongue complexity in the relatively narrow age range studied, age was negatively associated with tongue complexity for /t/ based on NINFL. For the comparison based on disorder status, the TD children were found to have more complex tongue shapes for /ɹ/ relative to the reference level, the vowel /æ/, based on NINFL. Finally, with diagnostic groups pooled, /ɹ/ productions that were perceived as accurate were found to have more complex tongue shapes than those perceived as inaccurate, also based on NINFL. Taken together, these results suggest that differences in tongue complexity are detectable across multiple developmental dimensions, a finding that supports the potential use of these measures as an index of motor control.

The third manuscript approached Aim 1 by investigating whether acoustically measured production accuracy of rhotic targets would increase as a function of tongue complexity in a sample of older children ages 9-14 undergoing treatment for RSE affecting rhotic targets. Based on MCI, tongue complexity for /ɹ/ was positively associated with the normalized distance between the second and third formant frequencies before treatment, but this relationship was reversed after treatment. Based on previous research showing a relationship between acoustically measured accuracy and expert listeners' perceptual ratings (Kabakoff et al., 2021; Preston et al., 2019), it is likely that at post-treatment, the more complex tongue shapes were also associated with greater perceived accuracy. This pattern suggested that the children acquired the changes in tongue shape from their experience in the ultrasound-based treatment program. Even though there was no control condition rule out the possibility of spontaneous

improvement, spontaneous improvement past the age of eight is considered unlikely for children with RSE (Gibbon & Paterson, 2006). Therefore, the present results suggest that treatment contributed to the higher tongue complexity and higher accuracy observed after treatment.

Aim 2: Determine whether there is a relationship between somatosensory acuity and tongue complexity

The third manuscript addressed this aim by asking whether an individual's degree of somatosensory acuity would predict degree of tongue complexity for rhotic targets in the sample of older children ages 9-14 with RSE affecting /ɹ/. Auditory acuity was controlled for in this analysis. At the pre-treatment time point, there was no significant association between tongue complexity (based on either metric) and somatosensory acuity or auditory acuity. However, at the post-treatment time point, there was a significant association between tongue complexity (as measured by MCI) and performance on both somatosensory and auditory acuity tasks. That is, the predicted relationships between sensory acuity and tongue complexity were not observed before treatment, but after treatment, the predicted pattern of stronger acuity being associated greater degrees of tongue complexity was observed in both auditory and somatosensory domains. While further investigation is needed, this pattern could suggest that those children with RSE who have strong acuity in one sensory domain are more able to achieve the higher complexity tongue shapes over the course of ultrasound-based biofeedback intervention.

Measuring sensorimotor skill

Ultrasound-based measures of tongue complexity

The rationale for developing measures of tongue shape complexity was to be able to distinguish children with isolated speech errors from those who have speech motor delay. Drawing from current models of speech production such as the DIVA model (Guenther, 2016), as children attempt execution of a range of motor plans to help them arrive at auditory speech goals, they gradually refine the specification of both auditory and somatosensory goals, which increases the proficiency of the feedforward control system in reaching sensory goals in a variety of coarticulatory contexts. We henceforth refer to the emergent robustness of the feedforward plan as "motor skill." Therefore, an overarching goal of this dissertation was to quantify individual differences in motor skill. Previous research has quantified motor skill using kinematic measures to determine the degree of coupling between articulators (Green et al., 2000) or the variability of movement trajectories (Goffman & Smith, 1999; Grigos, 2009; Terband et al., 2011); a separate line of research has used palatography and EPG to measure degree of lingual differentiation indirectly via linguopalatal contact patterns (Fletcher, 1989; Gibbon, 1999; Gibbon, Stewart, Hardcastle, & Crampin, 1999). However, direct measures of tongue shape may be optimally suited for quantifying degree of lingual differentiation because they take the entire tongue surface into consideration and do not rely on palatal contact, which is minimal for sonorants. Instead, the present study used ultrasound to measure tongue complexity in children as an index of motor skill. Results from the manuscripts

reported here suggest that ultrasound-based measures of tongue complexity are useful at distinguishing rhotic tongue shapes from TD children versus children with SSD and also in distinguishing correct from incorrect rhotic productions across diagnostic groups. Thus, the first contribution of this dissertation to basic science is the replication of findings from Preston et al. (2019), which validates the use ultrasound-based measures of tongue complexity as an accessible and child-friendly approach for quantifying motor skill in children.

The two metrics of tongue complexity that were considered throughout this dissertation were MCI and NINFL. (A Procrustes metric was also considered in the first manuscript, but was ruled out as a measure of tongue complexity on the basis of its reduced comparability between individuals.) The first manuscript revealed a mix of converging and diverging patterns for these two measures across phonemes. In the other two manuscripts, a single measure was responsible for the majority of significant findings, but the preferred measure differed across the two studies. That is, for younger children in the second manuscript, NINFL was associated with significant age-based differences in tongue shape for /t/ relative to /æ/, as well as significant differences based on diagnostic category and degree of perceived accuracy for rhotic targets. Comparable patterns were not apparent when MCI was used as the measure of tongue complexity. By contrast, for the older children with RSE in the third manuscript, MCI was a significant predictor of acoustically measured accuracy, and also participated in significant interactions between sensory acuity (in both auditory and somatosensory domains) and treatment time point. Equivalent effects were not observed in

models using NINFL as the measure of tongue complexity. As both MCI and NINFL are considered relatively robust across differences in vocal tract size, it is unlikely that the discrepancies in findings between the two measures across the two manuscripts are attributable to the differing ages of their participants. Instead, a possible explanation is suggested by the qualitative observation that the younger children in the second manuscript tended to produce more bunched than retroflex tongue shapes, whereas the older children receiving treatment in the third manuscript tended to use retroflex tongue shapes for /ɹ/. (It is possible that this difference represents a generalization about tongue shapes acquired over the course of ultrasound biofeedback training versus naturalistic learning, but further study would be needed to make such an assertion.)

Previous findings suggest that computational differences between MCI and NINFL could lead to differing results for retroflex versus bunched tongue shapes (Heyne et al., 2020; Preston et al., 2019; Stolar & Gick, 2013). Recall that MCI is driven by curvature while NINFL is determined by the number of changes of a shape from convex to concave (and vice versa). As such, tongue shapes with high amounts of local curvature (i.e., the radius of curvature is small at a given point) could yield high MCI values but low NINFL values.[9] If the younger children tended to produce rhotics with bunched tongue shapes, which have relatively high global but relatively low local curvature, their accurate productions

[9] However, this account is not consistent with recent work from Heyne et al. (2020), which found that higher MCI values were associated with tip-down variants in adults.

may have been associated with high NINFL but low MCI values, driving significant results in the analyses using NINFL only. By contrast, if the older children tended to produce rhotic targets with retroflex tongue shapes, which have a tongue tip constriction characterized by a high degree of local curvature, their accurate productions may have been characterized by high MCI values but relatively low NINFL values. While at the present time, this is admittedly a speculative explanation, efforts are underway to systematically code tongue contours from the current studies as retroflex versus bunched. The current findings suggest that computing both MCI and NINFL is advisable, particularly if the data under investigation are expected to contain a mix of retroflex and bunched tongue shapes for /ɹ/.

It is essential to acknowledge the limitation that the measures that were used to quantify degree of lingual differentiation as an index of motor skill across the three studies did not have optimal interrater reliability. For the young child sample used in the first and second manuscripts, reliability was moderate for MCI and fair for NINFL; for the older child sample used in the third manuscript, reliability was initially poor for MCI and fair for NINFL. In the analyses for both child samples, these differences were judged to be associated with specific decisions that were made by the individuals who were manually tracing the tongue shapes. This claim was corroborated by the improvement in reliability observed for the young child sample when manually-traced tongue shapes were replaced with contours generated by an automated tracing algorithm. For the older children, the relatively low reliability observed for MCI was judged to be

attributable to a specific decision on how far to trace the tongue tip anteriorly. Re-training students to conform to the conventions established in the original data set was crucial in ensuring adequate reliability, particularly for MCI. As discussed in the third manuscript, we attributed the consistently suboptimal reliability for NINFL to the sensitivity of the filter cutoff and threshold settings for this metric. We therefore caution researchers against using these metrics without careful optimization of tracing conventions (even in the context of automated tracing) as well as close attention to the calibration of the NINFL metric.

Somatosensory acuity

The rationale for developing measurements of somatosensory acuity was to be able to identify children with reduced access and/or ability to respond to somatosensory feedback. The DIVA model posits that skilled speech production involves translation of stored mental representations into somatosensory targets in combination with somatosensory feedback loops that direct articulator placement. The present study aimed to measure somatosensory acuity based on evidence that the specificity of an individual's somatosensory targets and access to somatosensory feedback can influence precision in speech production (Ghosh et al., 2010). At the same time, we controlled for the parallel channel of auditory acuity, which has been more thoroughly studied and is known to influence production outcomes across a wide range of contexts. Results from the third manuscript suggest that an oral stereognosis task can be used as an index of the tactile aspect of somatosensory acuity, and that somatosensory acuity was significantly associated with tongue shape complexity in children with RSE after

treatment targeting rhotics. A parallel pattern was observed in the auditory domain. Thus, the second contribution of this dissertation to basic science is that it corroborates past research indicating that somatosensory skill can be quantified in individual children, while offering the novel contribution that the quantification of this skill may, in turn, predict motor skill in individual children. However, in the present study the predicted relationship between somatosensory acuity and tongue complexity was only observed after participants completed a program of ultrasound-based biofeedback treatment, suggesting that sensory acuity may be more closely related to treatment response than to raw measures of motor skill. While the present study provided promising results from an oral stereognosis task, future research should compare various measures of somatosensory acuity, including those intended to measure the proprioceptive aspect, to determine whether another task may optimally quantify this skill in individuals varying in age in both typical and disordered populations.

Clinical implications

In addition to the basic science considerations discussed above, the present research also has translational considerations. The first translational contribution of this research is in proposing ultrasound measurement of tongue complexity as a method for assessing an individual's degree of motor skill. Because speech errors of unknown motor origin are less likely to resolve than isolated speech errors (Vick et al., 2014), it is important to be able to identify children with motor involvement in a timely manner. Accurate and readily accessible identification of the subset of children with SSD who also have reduced motor skill would present

the opportunity to match these children with a corresponding motor-based treatment approach. Results from all three manuscripts in this dissertation support the use of ultrasound-based measures of tongue complexity to measure motor skill, particularly for late-developing targets such as /l/ and /ɹ/. In the second manuscript, younger children with SSD were observed to produce rhotic targets with tongue shapes that were less complex than those of TD children. The reduced tongue complexity observed for late-developing targets in the children with SSD suggests that this group may be characterized by reduced motor skill. However, our results from the same manuscript did not support the presence of reduced tongue complexity for all targets (i.e., widespread undifferentiated gestures) in the younger children. The lack of evidence of covert error associated with widespread undifferentiated gestures represents a limitation of the present findings in terms of clinical applicability. That is, except in a few outlier productions in which tongue complexity does not correspond with accuracy, most of the present significant findings could in principle be reproduced using perceived accuracy instead of tongue complexity. Further study is needed to determine whether tongue complexity can sometimes reveal covert patterns in the populations studied here.

Despite the limited clinical application described above, the present data support the possibility that exceptions (in which tongue complexity does not agree with perceived accuracy) should be examined in more detail. That is, opposite from what was predicted, higher tongue complexity was associated with incorrect rhotic production for the older children before treatment in the third manuscript.

This pattern is compatible with the potential interpretation that before enrolling in our ultrasound-based biofeedback treatment study, the older children with RSE had already acquired complex tongue shapes as a result of prior treatment that did not result in the attainment of an adultlike rhotic percept. This suggests that these older children may have developed maladaptive patterns as a result of prior treatment, a phenomenon that would not be evident by measuring perceived accuracy on its own. Though speculative, this specific unexpected finding may lend support to the claim that tongue complexity offers clinically-relevant information beyond what perceived accuracy provides.

A final translational contribution of this research is that it offers data on an oral stereognosis task as a means to assess a child's degree of somatosensory acuity. Because somatosensory feedback is crucial for accurate speech, it follows that being able to measure this skill may be helpful in determining who has reduced access or ability to respond to this form of feedback. The results of the third manuscript showed a significant correlation in the predicted direction between somatosensory acuity and tongue complexity at the post-treatment time point in older children with RSE. Although more research is needed to explore any treatment-based effects, this finding suggests that children with strong somatosensory acuity may have derived more benefit from ultrasound-based treatment than those with weak somatosensory acuity. With further study, it may be possible to identify a recommended treatment approach based on a child's performance on somatosensory measures like the one reported here.

Future directions

The current dissertation provides the foundation for a line of research that measures motor and sensory abilities in children with the long-term goal of matching individual children with the treatment approach best aligned with their sensorimotor profile. The data from the younger children with and without SSD in the first and second manuscripts were instrumental in revealing phoneme-specific patterns of tongue complexity and differences in tongue complexity based on disorder status and degree of perceived accuracy. A next step is to compare tongue shape patterns in the subset of young children with SSD in the second manuscript (n = 3) who met criteria to participate in an ongoing study tracking changes in tongue shape over the course of phonological intervention. Another future direction is to compare tongue shapes between the younger and older children with SSD, as well as to collect a new sample of older TD children to compare tongue shape complexity with the current sample of older children with RSE. Collection of this sample will also enable the exploration of typical associations between sensory skills and speech outcomes for rhotic targets. McAllister Byun and Tiede (2017) reported the relationship between auditory acuity and speech production skill in TD children producing rhotic targets, but there is no comparable research examining somatosensory acuity in relation to rhotic production. The proposed new data collection will make it possible to examine both auditory and somatosensory acuity in connection with speech production skill in TD children producing rhotic targets. Finally, the sample of older TD children will also help determine whether the patterns observed at the

post-treatment time point in the third manuscript are also found in typical development. In the somatosensory domain, it will be important to examine other measures of somatosensory acuity, including tasks tapping proprioceptive as well as tactile aspects of somatosensation, to identify the best predictors of speech outcomes. Finally, for the ultimate purpose of being able to match individuals to the best treatment approach based on their sensorimotor profile, future research should measure these multiple dimensions of skill in children with SSD and examine them in relation to response to various forms of treatment, including ultrasound biofeedback and visual-acoustic biofeedback. The ability to pair children with an appropriate treatment approach has the potential to accelerate the treatment process, potentially contributing to a reduction in the number of individuals with RSE who persist with errors into adulthood.

BIBLIOGRAPHY

Abakarova, D., Iskarous, K., & Noiray, A. (2020). *Articulatory strategies and coarticulatory patterns across age.* Paper presented at the 12th International Seminar on Speech Production, Online.

Attanasio, J. S. (1987). Relationships between oral sensory feedback skills and adaptation to delayed auditory feedback. *Journal of Communication Disorders, 20*(5), 391-402. https://doi.org/10.1016/0021-9924(87)90027-X

Bacsfalvi, P. (2010). Attaining the lingual components of /r/ with ultrasound for three adolescents with cochlear implants. *Revue canadienne d'orthophonie et d'audiologie, 34*(3), 206.

Barlow, J. A., & Gierut, J. A. (2002). Minimal pair approaches to phonological remediation. *Seminars in Speech and Language, 23*(01), 57-68. https://doi.org/10.1055/s-2002-24969

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using 'lme4'. *Journal of Statistical Software, 67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Baum, S. R., & McFarland, D. H. (1997). The development of speech adaptation to an artificial palate. *The Journal of the Acoustical Society of America, 102*(4), 2353-2359. https://doi.org/10.1121/1.419619

Bernhardt, B., Bacsfalvi, P., Gick, B., Radanov, B., & Williams, R. (2005). Exploring the use of electropalatography and ultrasound in speech habilitation. *Journal of Speech-Language Pathology and Audiology, 29*(4), 169-182.

Bernhardt, B., Gick, B., Bacsfalvi, P., & Adler-Bock, M. (2005). Ultrasound in speech therapy with adolescents and adults. *Clinical Linguistics & Phonetics, 19*(6-7), 605-617. https://doi.org/10.1080/02699200500114028

Bernhardt, B., Gick, B., Bacsfalvi, P., & Ashdown, J. (2003). Speech habilitation of hard of hearing adolescents using electropalatography and ultrasound as evaluated by trained listeners. *Clinical Linguistics & Phonetics, 17*(3), 199-216. https://doi.org/10.1080/0269920031000071451

Berryman, L. J., Yau, J. M., & Hsiao, S. S. (2006). Representation of object size in the somatosensory system. *Journal of Neurophysiology, 96*(1), 27-39. https://doi.org/10.1152/jn.01190.2005

Boersma, P., & Hayes, B. (2001). Empirical tests of the gradual learning algorithm. *Linguistic inquiry, 32*(1), 45-86. https://doi.org/10.1162/002438901554586

Boersma, P., & Weenink, D. (2019). Praat: doing phonetics by computer (Version 6.0.50) [Computer program]. Retrieved from www.fon.hum.uva.nl/praat/

Borden, G. J., Harris, K. S., & Catena, L. (1973). Oral feedback II. An electromyographic study of speech under nerve-block anesthesia. *Journal of Phonetics, 1*(4), 297-308. https://doi.org/10.1016/S0095-4470(19)31399-3

Bosma, J. F. (1963). Maturation of function of the oral and pharyngeal region. *American Journal of Orthodontics and Dentofacial Orthopedics, 49*(2), 94-104. https://doi.org/10.1016/0002-9416(63)90097-6

Boyce, S. E. (2015). The articulatory phonetics of /r/ for residual speech errors. *Seminars in Speech and Language, 36*(04), 257-270. https://doi.org/10.1055/s-0035-1562909

Brunner, J., Fuchs, S., & Perrier, P. (2009). On the relationship between palate shape and articulatory behavior. *The Journal of the Acoustical Society of America, 125*(6), 3936-3949. https://doi.org/10.1121/1.3125313

Cabbage, K. L., Hogan, T. P., & Carrell, T. D. (2016). Speech perception differences in children with dyslexia and persistent speech delay. *Speech Communication, 82*, 14-25. https://doi.org/10.1016/j.specom.2016.05.002

Campbell, H., Harel, D., Hitchcock, E., & McAllister Byun, T. (2018). Selecting an acoustic correlate for automated measurement of American English rhotic production in children. *International Journal of Speech-Language Pathology, 20*(6), 635-643. https://doi.org/10.1080/17549507.2017.1359334

Campbell, T. F., Dollaghan, C. A., Rockette, H. E., Paradise, J. L., Feldman, H. M., Shriberg, L. D., Sabo, D. L., & Kurs-Lasky, M. (2003). Risk factors for speech delay of unknown origin in 3-year-old children. *Child Development, 74*(2), 346-357. https://doi.org/10.1111/1467-8624.7402002

Christensen, R. H. B. (2015). 'ordinal' package in R: Regression models for ordinal data via cumulative link (mixed) models (Version 2015.1-21). Retrieved from https://cran.r-project.org/web/packages/ordinal/index.html

Cialdella, L., Kabakoff, H., Preston, J. L., Dugan, S., Spencer, C., Boyce, S., Tiede, M., Whalen, D. H., & McAllister, T. (2021). Auditory-perceptual acuity in rhotic misarticulation: baseline characteristics and treatment response. *Clinical Linguistics & Phonetics, 35*(1), 19-42. https://doi.org/10.1080/02699206.2020.1739749

Cleland, J., Scobbie, J. M., Heyde, C., Roxburgh, Z., & Wrench, A. A. (2017). Covert contrast and covert errors in persistent velar fronting. *Clinical Linguistics & Phonetics, 31*(1), 35-55. https://doi.org/10.1080/02699206.2016.1209788

Cleland, J., Scobbie, J. M., & Wrench, A. A. (2015). Using ultrasound visual biofeedback to treat persistent primary speech sound disorders. *Clinical Linguistics & Phonetics, 29*(8-10), 575-597. https://doi.org/10.3109/02699206.2015.1016188

Crowe, K., & McLeod, S. (2020). Children's English consonant acquisition in the United States: A review. *American Journal of Speech-Language Pathology*, 1-15. https://doi.org/10.1044/2020_AJSLP-19-00168

Dawson, K. M. (2016). tshape_analysis (Version 3) [Python script]. Retrieved from https://github.com/kdawson2/tshape_analysis

Dawson, K. M., Tiede, M., & Whalen, D. H. (2016). Methods for quantifying tongue shape and complexity using ultrasound imaging. *Clinical Linguistics & Phonetics, 30*(3-5), 328-344. https://doi.org/10.3109/02699206.2015.1099164

Delattre, P., & Freeman, D. C. (1968). A dialect study of American r's by x-ray motion picture. *Linguistics, 6*(44), 29-68. https://doi.org/10.1515/ling.1968.6.44.29

Derrick, D., Carignan, C., Chen, W., Shujau, M., & Best, C. T. (2018). Three-dimensional printable ultrasound transducer stabilization system. *Journal of the Acoustical Society of America, 144*(5), EL392-EL398. https://doi.org/10.1121/1.5066350

Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test - 4th edition*. Bloomington, MN: Pearson Assessments.

Etter, N. M., Miller, O. M., & Ballard, K. J. (2017). Clinically available assessment measures for lingual and labial somatosensation in healthy adults: Normative data and test reliability. *American Journal of Speech-Language Pathology, 26*(3), 982-990. https://doi.org/10.1044/2017_AJSLP-16-0151

Felsenfeld, S., Broen, P. A., & McGue, M. (1994). A 28-year follow-up of adults with a history of moderate phonological disorder: educational and occupational results. *Journal of Speech, Language, and Hearing Research, 37*(6), 1341-1353. https://doi.org/10.1044/jshr.3706.1341

Feng, Y., Gracco, V. L., & Max, L. (2011). Integration of auditory and somatosensory error signals in the neural control of speech movements. *Journal of Neurophysiology, 106*(2), 667-679. https://doi.org/10.1152/jn.00638.2010

Fey, M. E. (1985). Articulation and phonology: Inextricable constructs in speech pathology. *Human Cammunication Canada/Communication Humaine Canada, 9*(1), 7-16.

Fletcher, S. G. (1989). Palatometric specification of stop, affricate, and sibilant sounds. *Journal of Speech, Language, and Hearing Research, 32*(4), 736-748. https://doi.org/10.1044/jshr.3204.736

Flipsen, P. (2015). Emergence and prevalence of persistent and residual speech errors. *Seminars in Speech and Language, 36*(4), 217-223. https://doi.org/10.1055/s-0035-1562905

Flipsen, P., Shriberg, L. D., Weismer, G., Karlsson, H. B., & McSweeny, J. L. (2001). Acoustic phenotypes for speech-genetics studies: reference data for residual /ɝ/ distortions. *Clinical Linguistics & Phonetics, 15*(8), 603-630. https://doi.org/10.1080/02699200110069410

Franken, M. K., Acheson, D. J., McQueen, J. M., Eisner, F., & Hagoort, P. (2017). Individual variability as a window on production-perception interactions in speech motor control. *The Journal of the Acoustical Society of America, 142*(4), 2007-2018. https://doi.org/10.1121/1.5006899

Fucci, D. J. (1972). Oral vibrotactile sensation: An evaluation of normal and defective speakers. *Journal of Speech and Hearing Research, 15*, 179-184. https://doi.org/10.1044/jshr.1501.179

Fucci, D. J., & Robertson, J. H. (1971). "Functional" defective articulation: An oral sensory disturbance. *Perceptual and Motor Skills, 33*(3), 711-714. https://doi.org/10.2466/pms.1971.33.3.711

Gammon, S. A., Smith, P. J., Daniloff, R. G., & Kim, C. W. (1971). Articulation and stress/juncture production under oral anesthetization and masking. *Journal of Speech and Hearing Research, 14*(2), 271-282. https://doi.org/10.1044/jshr.1402.271

Ghosh, S. S., Matthies, M. L., Maas, E., Hanson, A., Tiede, M., Ménard, L., Guenther, F. H., Lane, H., & Perkell, J. S. (2010). An investigation of the relation between sibilant production and somatosensory and auditory acuity. *Journal of the Acoustical Society of America, 128*(5), 3079-3087. https://doi.org/10.1121/1.3493430

Gibbon, F. E. (1999). Undifferentiated lingual gestures in children with articulation/phonological disorders. *Journal of Speech, Language, and Hearing Research, 42*(2), 382-397. https://doi.org/10.1044/jslhr.4202.382

Gibbon, F. E., Dent, H., & Hardcastle, W. (1993). Diagnosis and therapy of abnormal alveolar stops in a speech-disordered child using electropalatography. *Clinical Linguistics & Phonetics, 7*(4), 247-267. https://doi.org/10.1080/02699209308985565

Gibbon, F. E., Hardcastle, B., & Dent, H. (1995). A study of obstruent sounds in school-age children with speech disorders using electropalatography. *International Journal of Language and Communication Disorders, 30*(2), 213-225. https://doi.org/10.3109/13682829509082532

Gibbon, F. E., & Paterson, L. (2006). A survey of speech and language therapists' views on electropalatography therapy outcomes in Scotland. *Child Language Teaching and Therapy, 22*(3), 275-292. https://doi.org/10.1191/0265659006ct308xx

Gibbon, F. E., Stewart, F., Hardcastle, W. J., & Crampin, L. (1999). Widening access to electropalatography for children with persistent sound system disorders. *American Journal of Speech-Language Pathology, 8*(4), 319-334. https://doi.org/10.1044/1058-0360.0804.319

Gick, B., Allen, B., Roewer-Després, F., & Stavness, I. (2017). Speaking tongues are actively braced. *Journal of Speech, Language, and Hearing Research, 60*(3), 494-506. https://doi.org/10.1044/2016_JSLHR-S-15-0141

Gick, B., Bacsfalvi, P., Bernhardt, B. M., Oh, S., Stolar, S., & Wilson, I. (2007). A motor differentiation model for liquid substitutions in children's speech. *Proceedings of Meetings on Acoustics, 1*, 060003. https://doi.org/10.1121/1.2951481

Goffman, L., & Smith, A. (1999). Development and phonetic differentiation of speech movement patterns. *Journal of Experimental Psychology: Human Perception and Performance, 25*(3), 649. https://doi.org/10.1037/0096-1523.25.3.649

Goldman, R., & Fristoe, M. (2000). *Goldman-Fristoe Test of Articulation - 2nd Edition (GFTA-2)* Bloomington, MN: Pearson/PsychCorp.

Goodall, C. (1991). Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society. Series B (Methodological)*, 285-339. https://doi.org/10.1111/j.2517-6161.1991.tb01825.x

Green, J. R., Moore, C. A., Higashikawa, M., & Steeve, R. W. (2000). The physiologic development of speech motor control: Lip and jaw coordination. *Journal of Speech, Language, and Hearing Research, 43*(1), 239-255. https://doi.org/10.1044/jslhr.4301.239

Grigos, M. I. (2009). Changes in articulator movement variability during phonemic development: a longitudinal study. *Journal of Speech,*

Language, and Hearing Research, 52(1), 164-177. https://doi.org/10.1044/1092-4388%282008/07-0220%29
Gritsyk, O., Kabakoff, H., Li, J. J., Ayala, S., Shiller, D. M., & McAllister, T. (2021). Toward an index of oral somatosensory acuity: Comparison of three measures in adults. *Perspectives of the ASHA Special Interest Groups, 6*(3), 500-512. https://doi.org/10.1044/2021_PERSP-20-00218
Guenther, F. H. (2016). *Neural Control of Speech*. Cambridge, MA: Massachusetts Institute of Technology.
Harel, D., & McAllister, T. (2019). Multilevel models for communication sciences and disorders. *Journal of Speech, Language, and Hearing Research, 62*(4), 783-801. https://doi.org/doi:10.1044/2018_JSLHR-S-18-0075
Hazan, V., & Barrett, S. (2000). The development of phonemic categorization in children aged 6–12. *Journal of Phonetics, 28*(4), 377-396. https://doi.org/10.1006/jpho.2000.0121
Hearnshaw, S., Baker, E., & Munro, N. (2018). The speech perception skills of children with and without speech sound disorder. *Journal of Communication Disorders, 71*, 61-71. https://doi.org/10.1016/j.jcomdis.2017.12.004
Hearnshaw, S., Baker, E., & Munro, N. (2019). Speech perception skills of children with speech sound disorders: A systematic review and meta-analysis. *Journal of Speech, Language, and Hearing Research, 62*(10), 3771-3789. https://doi.org/10.1044/2019_JSLHR-S-18-0519
Hedlund, G., & Rose, Y. (2020). Phon (Version 3.2.1-beta.1) [Computer Software]. Retrieved from https://phon.ca
Heyne, M., Wang, X., Derrick, D., Dorreen, K., & Watson, K. (2020). The articulation of /ɹ/ in New Zealand English. *Journal of the International Phonetic Association, 50*(3), 366-388. https://doi.org/10.1017/S0025100318000324
Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews: Neuroscience, 13*(2), 135-145. https://doi.org/10.1038/nrn3158
Hitchcock, E., Harel, D., & McAllister Byun, T. (2015). Social, emotional, and academic impact of residual speech errors in school-aged children: A survey study. *Seminars in Speech and Language, 36*(4), 283-293. https://doi.org/10.1055/s-0035-1562911
Hodson, B. W. (2004). *Hodson Assessment of Phonological Patterns, 3rd ed.* Austin, TX: Pro-Ed.
Hodson, B. W., & Paden, E. P. (1983). *Targeting Intelligible Speech: A Phonological Approach to Remediation*. San Diego, CA: College-Hill Press.
Houde, J. F., & Nagarajan, S. S. (2011). Speech production as state feedback control. *Frontiers in Human Neuroscience, 5*, 82. https://doi.org/10.3389/fnhum.2011.00082
Ibrahim, J. G., & Molenberghs, G. (2009). Missing data methods in longitudinal studies: a review. *Test (Madrid, Spain), 18*(1), 1-43. https://doi.org/10.1007/s11749-009-0138-x

Jamieson, D. G., & Rvachew, S. (1992). Remediating speech production errors with sound identification training. *Journal of Speech-Language Pathology and Audiology, 16*(3), 201-210. https://doi.org/1993-27184-001

Kabakoff, H., Beames, S. P., Tiede, M., Whalen, D. H., & McAllister, T. (under review). Comparing metrics for quantification of children's tongue shape complexity using ultrasound imaging. https://doi.org/10.17605/OSF.IO/NE2VS

Kabakoff, H., Gritsyk, O., Harel, D., Tiede, M., Preston, J. L., Whalen, D. H., & McAllister, T. (submitted). Characterizing sensorimotor profiles in children with residual speech errors: a pilot study. https://doi.org/10.17605/OSF.IO/78ZQB

Kabakoff, H., Gritsyk, O., Li, J. J., Ayala, S., Hitchcock, E., Preston, J. L., Harel, D., Shiller, D. M., & McAllister, T. (2020). Tracking development of somatosensory acuity: Age-based comparison of three measures. In *12th International Seminar on Speech Perception*. Online.

Kabakoff, H., Harel, D., Tiede, M., Whalen, D. H., & McAllister, T. (2021). Extending ultrasound tongue complexity measures to speech development and disorders. *Journal of Speech, Language, and Hearing Research, 64*(7), 2557-2574. https://doi.org/10.1044/2021_JSLHR-20-00537

Kent, R. D. (1992). The biology of phonological development. In C. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological development: Models, research, implications* (pp. 65-90). Timonium, MD: York Press, Inc.

Klein, H. B., McAllister Byun, T., Davidson, L., & Grigos, M. I. (2013). A multidimensional investigation of children's /r/ productions: perceptual, ultrasound, and acoustic measures. *American Journal of Speech-Language Pathology, 22*(3), 540-553. https://doi.org/10.1044/1058-0360%282013/12-0137%29

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Lametti, D. R., Nasir, S. M., & Ostry, D. J. (2012). Sensory preference in speech production revealed by simultaneous alteration of auditory and somatosensory feedback. *Journal of Neuroscience, 32*(27), 9351-9358. https://doi.org/10.1523/JNEUROSCI.0404-12.2012

Laporte, C., & Ménard, L. (2018). Multi-hypothesis tracking of the tongue surface in ultrasound video recordings of normal and impaired speech. *Medical Image Analysis, 44*, 98-114. https://doi.org/10.1016/j.media.2017.12.003

Lee, A., Gibbon, F. E., & O'Donovan, C. (2013). Tongue-palate contact of perceptually acceptable alveolar stops. *Clinical Linguistics & Phonetics, 27*(4), 312-321. https://doi.org/10.3109/02699206.2012.757651

Lee, S., Potamianos, A., & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *Journal of the Acoustical Society of America, 105*(3), 1455-1468. https://doi.org/10.1121/1.426686

Lennes, M. (2003). Collect formant data from files [Praat script]. Retrieved from https://lennes.github.io/spect/scripts/collect_formant_data_from_files.praat

Li, M., Kambhamettu, C., & Stone, M. (2005). Automatic contour tracking in ultrasound images. *Clinical Linguistics & Phonetics, 19*(6-7), 545-554. https://doi.org/10.1080/02699200500113616

Liégeois, F. J., & Morgan, A. T. (2012). Neural bases of childhood speech disorders: lateralization and plasticity for speech functions during development. *Neuroscience and Biobehavioral Reviews, 36*(1), 439-458. https://doi.org/10.1016/j.neubiorev.2011.07.011

Liljencrants, J. (1971). Fourier series description of the tongue profile. *Speech Transmission Laboratory-Quarterly Progress Status Reports, 12*(4), 9-18.

Lin, S., & Demuth, K. (2015). Children's acquisition of English onset and coda /l/: Articulatory evidence. *Journal of Speech, Language, and Hearing Research, 58*(1), 13-27. https://doi.org/10.1044/2014_JSLHR-S-14-0041

Lulich, S. M., & Pearson, W. G., Jr. (2019). Three-/four-dimensional ultrasound technology in speech research. *Perspectives of the ASHA Special Interest Groups, 4*, 733-747. https://doi.org/10.1044/2019_PERS-SIG19-2019-0001

Maas, E., Mailend, M.-L., & Guenther, F. H. (2015). Feedforward and feedback control in apraxia of speech (AOS): Effects of noise masking on vowel production. *Journal of Speech, Language, and Hearing Research*. https://doi.org/10.1044/2014_JSLHR-S-13-0300

Maas, E., Robin, D. A., Hula, S. N. A., Freedman, S. E., Wulf, G., Ballard, K. J., & Schmidt, R. A. (2008). Principles of motor learning in treatment of motor speech disorders. *American Journal of Speech-Language Pathology, 17*(3), 277-298. https://doi.org/10.1044/1058-0360%282008/025%29

Macken, M. A., & Barton, D. (1980). The acquisition of the voicing contrast in English: A study of voice onset time in word-initial stop consonants. *Journal of Child Language, 7*(1), 41-74. https://doi.org/10.1017/S0305000900007029

MathWorks Inc. (2000). Matlab (Version 6.1) [Computer program]. Natick, MA. Retrieved from https://www.mathworks.com/products/matlab/

McAllister Byun, T., Buchwald, A., & Mizoguchi, A. (2016). Covert contrast in velar fronting: An acoustic and ultrasound study. *Clinical Linguistics & Phonetics, 30*(3-5), 249-276. https://doi.org/10.3109/02699206.2015.1056884

McAllister Byun, T., & Rose, Y. (2016). Analyzing clinical phonological data using Phon. *Seminars in Speech and Language, 37*(2), 85-105. https://doi.org/10.1055/s-0036-1580741

McAllister Byun, T., & Tessier, A. M. (2016). Motor influences on grammar in an emergentist model of phonology. *Language and Linguistics Compass, 10*(9), 431-452. https://doi.org/10.1111/lnc3.12205

McAllister Byun, T., & Tiede, M. (2017). Perception-production relations in later development of American English rhotics. *PloS One, 12*(2), e0172022. https://doi.org/10.1371/journal.pone.0172022

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica, 22*(3), 276-282. https://doi.org/10.11613/BM.2012.031

McLeod, S., & Crowe, K. (2018). Children's consonant acquisition in 27 languages: A cross-linguistic review. *American Journal of Speech-Language Pathology, 27*(4), 1546-1571. https://doi.org/10.1044/2018_AJSLP-17-0100

McNutt, J. C. (1977). Oral sensory and motor behaviors of children with /s/ or /r/ misarticulations. *Journal of Speech, Language, and Hearing Research, 20*(4), 694-703. https://doi.org/10.1044/jshr.2101.192

McReynolds, L. V., & Bennett, S. (1972). Distinctive feature generalization in articulation training. *Journal of Speech and Hearing Disorders, 37*(4), 462-470. https://doi.org/10.1044/jshd.3704.462

Ménard, L., Aubin, J., Thibeault, M., & Richard, G. (2012). Measuring tongue shapes and positions with ultrasound imaging: A validation experiment using an articulatory model. *Folia Phoniatrica et Logopaedica, 64*(2), 64-72. https://doi.org/10.1159/000331997

Mielke, J., Baker, A., & Archangeli, D. (2016). Individual-level contact limits phonological complexity: Evidence from bunched and retroflex /ɹ/. *Language, 92*(1), 101-140. https://doi.org/10.1353/lan.2016.0019

Murray, E., McCabe, P., Heard, R., & Ballard, K. J. (2015). Differential diagnosis of children with suspected childhood apraxia of speech. *Journal of Speech, Language, and Hearing Research, 58*(1), 43-60. https://doi.org/10.1044/2014_JSLHR-S-12-0358

Namasivayam, A. K., Coleman, D., O'Dwyer, A., & van Lieshout, P. (2020). Speech sound disorders in children: An articulatory phonology perspective. *Frontiers in Psychology, 10*(2998). https://doi.org/10.3389/fpsyg.2019.02998

Nasir, S. M., & Ostry, D. J. (2008). Speech motor learning in profoundly deaf adults. *Nature Neuroscience, 11*(10), 1217. https://doi.org/10.1038/nn.2193

Noiray, A., Abakarova, D., Rubertus, E., Krüger, S., & Tiede, M. (2018). How do children organize their speech in the first years of life? Insight from ultrasound imaging. *Journal of Speech, Language, and Hearing Research, 61*(6), 1355-1368. https://doi.org/10.1044/2018_JSLHR-S-17-0148

Noiray, A., Ménard, L., & Iskarous, K. (2013). The development of motor synergies in children: Ultrasound and acoustic measurements. *The Journal of the Acoustical Society of America, 133*(1), 444-452. https://doi.org/10.1121/1.4763983

Parrell, B., Ramanarayanan, V., Nagarajan, S., & Houde, J. (2019). The FACTS model of speech motor control: Fusing state estimation and task-based control. *PLoS Computational Biology, 15*(9), e1007321. https://doi.org/10.1371/journal.pcbi.1007321

Perkell, J. S., Matthies, M. L., Tiede, M., Lane, H., Zandipour, M., Marrone, N., Stockmann, E., & Guenther, F. H. (2004). The distinctness of speakers' /s/-/ʃ/ contrast is related to their auditory discrimination and use of an articulatory saturation effect. *Journal of Speech, Language, and Hearing Research, 47*(6), 1259-1269. https://doi.org/10.1044/1092-4388%282004/095%29

Perrachione, T. K., Lee, J., Ha, L. Y., & Wong, P. C. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *Journal of the Acoustical Society of America, 130*(1), 461-472. https://doi.org/10.1121/1.3593366

Preston, J. L., Hitchcock, E. R., & Leece, M. C. (2020). Auditory perception and ultrasound biofeedback treatment outcomes for children with residual /ɹ/ distortions: A randomized controlled trial. *Journal of Speech, Language, and Hearing Research, 63*(2), 444-455. https://doi.org/10.1044/2019_JSLHR-19-00060

Preston, J. L., McCabe, P., Tiede, M., & Whalen, D. H. (2019). Tongue shapes for rhotics in school-age children with and without residual speech errors. *Clinical Linguistics & Phonetics, 33*(4), 334-348. https://doi.org/10.1080/02699206.2018.1517190

Putnam, A. H., & Ringel, R. L. (1976). A cineradiographic study of articulation in two talkers with temporarily induced oral sensory deprivation. *Journal of Speech and Hearing Research, 19*(2), 247-266. https://doi.org/10.1044/jshr.1902.247

Python Software Foundation. (2016). Python programming language (Version 3.5.2) [Computer program]. Retrieved from https://www.python.org/downloads/

R Core Team. (2019). R: A language and environment for statistical computing [Programming language]: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Ringel, R. L., Burk, K. W., & Scott, C. M. (1968). Tactile perception: Form discrimination in the mouth. *British Journal of Disorders of Communication, 3*(2), 150-155. https://doi.org/10.3109/13682826809011454

Ringel, R. L., & Steer, M. D. (1963). Some effects of tactile and auditory alterations on speech output. *Journal of Speech and Hearing Research, 6*(4), 369-378. https://doi.org/10.1044/jshr.0604.369

RStudio Team. (2019). RStudio: integrated development for R (Version 1.2.1335) [Computer program]. Boston, MA: RStudio, Inc. Retrieved from https://www.rstudio.com/products/rstudio/

Ruscello, D. (1995). Client response to treatment in schools: A survey. *Unpublished raw data*.

Rvachew, S. (1994). Speech perception training can facilitate sound production learning. *Journal of Speech, Language, and Hearing Research, 37*(2), 347-357. https://doi.org/10.1044/jshr.3702.347

Rvachew, S., & Brosseau-Lapré, F. (2010). Speech perception intervention. In A. L. Williams, S. McLeod, & R. J. McCauley (Eds.), *Interventions for speech sound disorders in children.* Baltimore, MD: Paul H. Brookes.

Rvachew, S., & Brosseau-Lapré, F. (2012). An input-focused intervention for children with developmental phonological disorders. *SIG 1 Perspectives on Language Learning and Education, 19*(1), 31-35. https://doi.org/10.1044/lle19.1.31

Rvachew, S., & Brosseau-Lapré, F. (2015). A randomized trial of 12-week interventions for the treatment of developmental phonological disorder in Francophone children. *American Journal of Speech-Language Pathology, 24*(4), 637-658. https://doi.org/10.1044/2015_AJSLP-14-0056

Rvachew, S., Hodge, M., & Ohberg, A. (2005). Obtaining and interpreting maximum performance tasks from children: A tutorial. *Journal of Speech Language Pathology and Audiology, 29*(4), 146.

Rvachew, S., & Jamieson, D. G. (1989). Perception of voiceless fricatives by children with a functional articulation disorder. *Journal of Speech and Hearing Disorders, 54*(2), 193-208. https://doi.org/10.1044/jshd.5402.193

Rvachew, S., Nowak, M., & Cloutier, G. (2004). Effect of phonemic perception training on the speech production and phonological awareness skills of children with expressive phonological delay. *American Journal of Speech-Language Pathology, 13*(3), 250-263. https://doi.org/10.1044/1058-0360%282004/026%29

Rvachew, S., Ohberg, A., Grawburg, M., & Heyding, J. (2003). Phonological awareness and phonemic perception in 4-year-old children with delayed expressive phonology skills. *American Journal of Speech-Language Pathology*(4), 463-471. https://doi.org/10.1044/1058-0360(2003/092)

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

Semel, E., Wiig, E. H., & Secord, W. A. (2013). *Clinical Evaluation of Language Fundamentals, 5th Ed. (CELF-5)*. Toronto Ontario: Pearson: The Psychological Corporation.

Shawker, T. H., & Sonies, B. C. (1985). Ultrasound biofeedback for speech training. Instrumentation and preliminary results. *Investigative Radiology, 20*(1), 90-93. https://doi.org/10.1097/00004424-198501000-00022

Shiller, D. M., Rvachew, S., & Brosseau-Lapré, F. (2010). Importance of the auditory perceptual target to the achievement of speech production accuracy. *Revue canadienne d'orthophonie et d'audiologie, 34*(3), 181.

Shriberg, L. D. (1993). Four new speech and prosody-voice measures for genetics research and other studies in developmental phonological disorders. *Journal of Speech, Language, and Hearing Research, 36*(1), 105-140. https://doi.org/10.1044/jshr.3601.105

Shriberg, L. D., Fourakis, M., Hall, S. D., Karlsson, H. B., Lohmeier, H. L., McSweeny, J. L., Potter, N. L., Scheer-Cohen, A. R., Strand, E. A., & Tilkens, C. M. (2010). Extensions to the speech disorders classification

system (SDCS). *Clinical Linguistics & Phonetics, 24*(10), 795-824. https://doi.org/10.3109/02699206.2010.503006

Shriberg, L. D., Kwiatkowski, J., & Mabie, H. L. (2019). Estimates of the prevalence of motor speech disorders in children with idiopathic speech delay. *Clinical Linguistics & Phonetics*, 1-28. https://doi.org/10.1080/02699206.2019.1595731

Shriberg, L. D., & Lohmeier, H. L. (2008). *The Syllable Repetition Task (SRT).* University of Wisconsin-Madison, Madison, WI. Retrieved from http://www2.waisman.wisc.edu/phonology/techreports/TREP14.pdf

Shriberg, L. D., Lohmeier, H. L., Campbell, T. F., Dollaghan, C. A., Green, J. R., & Moore, C. A. (2009). A nonword repetition task for speakers with misarticulations: The Syllable Repetition Task (SRT). *Journal of Speech, Language, and Hearing Research, 52*, 1189-1212. https://doi.org/10.1044/1092-4388%282009/08-0047%29

Shriberg, L. D., Lohmeier, H. L., Strand, E., & Jakielski, K. J. (2012). Encoding, memory, and transcoding deficits in Childhood Apraxia of Speech. *Clin Linguist Phon, 26*(5), 445-482. https://doi.org/10.3109/02699206.2012.655841

Shriberg, L. D., Tomblin, J. B., & McSweeny, J. L. (1999). Prevalence of speech delay in 6-year-old children and comorbidity with language impairment. *Journal of Speech, Language, and Hearing Research, 42*(6), 1461-1481. https://doi.org/10.1044/jslhr.4206.1461

Smit, A. B., Hand, L., Freilinger, J. J., Bernthal, J. E., & Bird, A. (1990). The Iowa articulation norms project and its Nebraska replication. *Journal of Speech and Hearing Disorders, 55*(4), 779-798. https://doi.org/10.1044/jshd.5504.779

Steele, C. M., Hill, L., Stokely, S., & Peladeau-Pigeon, M. (2014). Age and strength influences on lingual tactile acuity. *Journal of texture studies, 45*(4), 317-323. https://doi.org/10.1111/jtxs.12076

Stolar, S., & Gick, B. (2013). An index for quantifying tongue curvature. *Canadian Acoustics, 41*(1), 11-15.

Stone, M. (2005). A guide to analysing tongue motion from ultrasound images. *Clinical Linguistics & Phonetics, 19*(6/7), 455-501. https://doi.org/10.1080/02699200500113558

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

Terband, H., Maassen, B., Guenther, F. H., & Brumberg, J. (2014). Auditory–motor interactions in pediatric motor speech disorders: Neurocomputational modeling of disordered development. *Journal of Communication Disorders, 47*, 17-33. https://doi.org/10.1016/j.jcomdis.2014.01.001

Terband, H., Maassen, B., Van Lieshout, P., & Nijland, L. (2011). Stability and composition of functional synergies for speech movements in children

with developmental speech disorders. *Journal of Communication Disorders, 44*(1), 59-74. https://doi.org/10.1016/j.jcomdis.2010.07.003
Tesar, B., & Smolensky, P. (1998). Learnability in optimality theory. *Linguistic inquiry, 29*(2), 229-268. https://doi.org/10.1162/002438998553734
Tiede, M. (2016). GetContours (Version 1.3) [Matlab script]. Retrieved from https://github.com/mktiede/GetContours
Tiede, M. (2020). GetContours (Version 2.4 UltraFest Edition) [Matlab script]. Retrieved from https://github.com/mktiede/GetContours
Tiede, M., Boyce, S. E., Holland, C. K., & Choe, K. A. (2004). A new taxonomy of American English /r/ using MRI and ultrasound. *Journal of the Acoustical Society of America, 115*(5), 2633-2634. https://doi.org/10.1121/1.4784878
Tremblay, S., Shiller, D. M., & Ostry, D. J. (2003). Somatosensory basis of speech production. *Nature, 423*(6942), 866-869. https://doi.org/10.1038/nature01710
Van Riper, C. (1978). *Speech correction: Principles and methods (6th edition)*. Englewood Cliffs, NJ: Prentice-Hall.
Vick, J. C., Campbell, T. F., Shriberg, L. D., Green, J. R., Truemper, K., Rusiewicz, H. L., & Moore, C. A. (2014). Data-driven subclassification of speech sound disorders in preschool children. *Journal of Speech, Language, and Hearing Research, 57*(6), 2033-2050. https://doi.org/10.1044/2014_JSLHR-S-12-0193
Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (1999). *Comprehensive Test of Phonological Processing- 2nd ed. (CTOPP-2)*: Pro-ed Austin, TX.
Whalen, D. H., Iskarous, K., Tiede, M., Ostry, D. J., Lehnert-LeHouillier, H., Vatikiotis-Bateson, E., & Hailey, D. S. (2005). The Haskins optically corrected ultrasound system (HOCUS). *Journal of Speech, Language, and Hearing Research, 48*(3), 543-553. https://doi.org/10.1044/1092-4388%282005/037%29
Wiig, E. H., Secord, W., & Semel, E. M. (2004). *Clinical Evaluation of Language Fundamentals Preschool, 2nd Ed.* Toronto Ontario: Pearson/PsychCorp.
Wong, P. C., & Perrachione, T. K. (2007). Learning pitch patterns in lexical identification by native English-speaking adults. *Applied Psycholinguistics, 28*(4), 565-585. https://doi.org/10.1017/s0142716407070312
Wong, P. C., Vuong, L. C., & Liu, K. (2017). Personalized learning: From neurogenetics of behaviors to designing optimal language training. *Neuropsychologia, 98*, 192-200. https://doi.org/10.1016/j.neuropsychologia.2016.10.002
Wren, Y., Miller, L. L., Peters, T. J., Emond, A., & Roulstone, S. (2016). Prevalence and predictors of persistent speech sound disorder at eight years old: Findings from a population cohort study. *Journal of Speech, Language, and Hearing Research, 59*(4), 647-673. https://doi.org/10.1044/2015_JSLHR-S-14-0282

Zandipour, M., Perkell, J. S., Guenther, F., Tiede, M., Honda, K., & Murano, E. (2006). Speaking with a bite-block: Data and modeling. In *Proceedings of the 7th International Seminar on Speech Production* (pp. 361-368). Ubatuba, Brazil: CEFALA.

Zharkova, N., Gibbon, F. E., & Hardcastle, W. J. (2015). Quantifying lingual coarticulation using ultrasound imaging data collected with and without head stabilisation. *Clinical Linguistics & Phonetics, 29*(4), 249-265. https://doi.org/10.3109/02699206.2015.1007528

APPENDICES

classification	site	subject	sex	age	CELF-P4 receptive tasks	PPVT-4	SRT PCC (%)	SRT Additions	SAILS (%)	HAPP-3	HAPP-3 identified candidate patterns
TD	Haskins	01F	F	5;6	103	x	x	x	86.7	97	*did not analyze candidate patterns for eligibility*
TD	Haskins	02F	F	5;11	100	x	x	x	71.7	95	*did not analyze candidate patterns for eligibility*
TD	Haskins	03M	M	5;6	111	x	x	x	71.7	85	*did not analyze candidate patterns for eligibility*
TD	Haskins	04F	F	5;6	113	x	x	x	86.7	97	*did not analyze candidate patterns for eligibility*
TD	Haskins	05M	M	4;2	121	x	x	x	68.3	90	*did not analyze candidate patterns for eligibility*
TD	Haskins	06M	M	4;11	142	x	x	x	81.7	87	*did not analyze candidate patterns for eligibility*
TD	Haskins	07F	F	5;11	111	x	x	x	83.3	88	*did not analyze candidate patterns for eligibility*
TD	Haskins	08F	F	5;9	95	x	x	x	75.0	88	*did not analyze candidate patterns for eligibility*
TD	Molloy	09F	F	4;3	103	x	x	x	74.6	112	*did not analyze candidate patterns for eligibility*
TD	Molloy	10F	F	6;0	117	x	x	x	76.7	85	*did not analyze candidate patterns for eligibility*
TD	Molloy	11F	F	4;9	100	x	x	x	53.3	81	*did not analyze candidate patterns for eligibility*
TD	Molloy	12M	M	6;3	105	x	x	x	90.0	100	*did not analyze candidate patterns for eligibility*
TD	Molloy	13F	F	4;3	112	x	x	x	70.0	103	*did not analyze candidate patterns for eligibility*
TD	Molloy	14F	F	4;5	125	x	x	x	85.0	100	*did not analyze candidate patterns for eligibility*
TD	Syracuse	15M	M	5;3	118	x	x	x	78.3	100	*did not analyze candidate patterns for eligibility*
TD	Syracuse	16F	F	5;5	115	x	x	x	70	84	*did not analyze candidate patterns for eligibility*
TD	Syracuse	17M	M	4;6	103	x	x	x	80	99	*did not analyze candidate patterns for eligibility*
SSD	Haskins	18M	M	4;2	127	136	68	3	70.0	<55	*consonant clusters, prevocalic liquids; stridents*
SSD	Haskins	19F	F	5;0	93	99	84	0	58.3	<55	*consonant clusters; prevocalic liquids; stridents*
SSD	Haskins	20M	M	5;4	98	113	88	x	75.00	<55	*consonant clusters, prevocalic liquids; velars*
SSD	Molloy	21M	M	4;0	90	x	x	x	71.7	74	*stridents; supplemental patterns: frontal lisp (67%); th/f (60%)*
SSD	Syracuse	22M	M	5;3	71	75	62	5	51.7	56	*prevocalic liquids; no supplemental patterns >40%*
SSD	Syracuse	23M	M	4;0	85	106	76	3	70.0	74	*consonant sequences/clusters; no supplemental patterns >40%*
SSD	Syracuse	24M	M	5;11	103	x	x	x	x	61	*prevocalic liquids; th substitutions (80%) identified on supplemental patterns worksheet; no 3rd pattern >40%*

Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's method ['lmerModLmerTest']

Formula: MCI ~ scale(age_mos, scale = F) * target_IPA + (1 | word) + (1 + target_IPA | subject)
Data: fuldat_TD

AIC	BIC	logLik	deviance	df.resid
1289.4	1494	-600.7	1201.4	728

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.3787	-0.679	-0.1384	0.5806	6.2247

Random effects:

Groups	Name	Variance	Std.Dev.	Corr					
subject	(Intercept)	0.046610563	0.2158948						
target_IPAw		0.049283873	0.2219997	-0.84					
target_IPAt		0.016539473	0.1286059	-0.22	0.58				
target_IPAk		0.013140982	0.1146341	-0.29	0.67	0.72			
target_IPAj		0.013168355	0.1147535	0.06	-0.31	-0.46	0.04		
target_IPAl		0.108144983	0.328854	-0.77	0.89	0.37	0.77	0.14	
target_IPAɹ		0.084438359	0.2905828	-0.34	0.62	0.7	0.28	-0.92	0.21
word	(Intercept)	0.000000129	0.0003594						
Residual		0.25506936	0.5050439						

Number of obs: 772, groups: subject, 17; word, 16

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	2.442003	0.067167	16.945151	36.357	0.0000000000000002***
scale(age_mos,scale=F)	-0.006541	0.008427	17.900938	-0.776	0.447735
target_IPAw	**0.218886**	**0.089134**	**23.021234**	**2.456**	**0.022032***
target_IPAt	-0.061809	0.068753	15.183661	-0.899	0.382687
target_IPAk	**0.300681**	**0.066517**	**25.854092**	**4.52**	**0.000121*****
target_IPAj	0.091522	0.096015	60.902563	0.953	0.344252
target_IPAl	0.173385	0.10838	17.003203	1.6	0.12806
target_IPAɹ	**0.430374**	**0.093457**	**17.023663**	**4.605**	**0.000251*****
scale(age_mos,scale=F):target_IPAw	0.005079	0.011454	25.261336	0.443	0.661202
scale(age_mos,scale=F):target_IPAt	-0.004923	0.0088	15.888447	-0.559	0.583686
scale(age_mos,scale=F):target_IPAk	-0.005559	0.008651	27.640997	-0.643	0.525799
scale(age_mos,scale=F):target_IPAj	0.00786	0.012968	79.823227	0.606	0.546168
scale(age_mos,scale=F):target_IPAl	0.003433	0.013514	17.340128	0.254	0.802439
scale(age_mos,scale=F):target_IPAɹ	0.015082	0.011686	17.977279	1.291	0.213188

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Cumulative Link Mixed Model fitted with the Laplace approximation

formula: NINFL ~ scale(age_mos, scale = F) * target_IPA + (1 | word) + (1 + target_IPA | subject)
data: fuldat_TD

link	threshold	nobs	logLik	AIC	niter	max.grad	cond.H
logit	flexible	772	-968.1	2028.19	9369(37434)	2.24E-03	4.00E+03

Random effects:

Groups	Name	Variance	Std.Dev.	Corr					
subject	(Intercept)	2.1313	1.4599						
target_IPAw		0.8418	0.9175	-0.478					
target_IPAt		0.1903	0.4363	-0.266	0.507				
target_IPAk		0.7749	0.8803	-0.68	0.852	0.387			
target_IPAj		0.6781	0.8235	-0.355	0.79	-0.121	0.665		
target_IPAl		1.6397	1.2805	-0.882	0.398	0.626	0.538	0.015	
target_IPAɹ		1.7977	1.3408	-0.877	0.494	0.408	0.452	0.316	0.864
word	(Intercept)	0	0						

Number of groups: subject 17, word 16

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
scale(age_mos,scale=F)	-0.004417	0.049752	-0.089	0.92926
target_IPAw	0.022284	0.383818	0.058	0.9537
target_IPAt	0.118134	0.284695	0.415	0.67818
target_IPAk	0.304924	0.333408	0.915	0.36042
target_IPAj	0.089685	0.435712	0.206	0.83692
target_IPAl	**0.881444**	**0.425303**	**2.073**	**0.03822***
target_IPAɹ	**1.270267**	**0.411853**	**3.084**	**0.00204****
scale(age_mos,scale=F):target_IPAw	0.034313	0.048428	0.709	0.47862
scale(age_mos,scale=F):target_IPAt	**-0.077731**	**0.035844**	**-2.169**	**0.03011***
scale(age_mos,scale=F):target_IPAt	-0.024905	0.042557	-0.585	0.55841
scale(age_mos,scale=F):target_IPAt	-0.03396	0.057163	-0.594	0.55246
scale(age_mos,scale=F):target_IPAl	-0.027915	0.052384	-0.533	0.5941
scale(age_mos,scale=F):target_IPAɹ	0.034009	0.050773	0.67	0.50297

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Threshold coefficients:

	Estimate	Std. Error	z value
1|2	-0.4175	0.4042	-1.033
2|3	1.073	0.4056	2.645
3|4	2.9925	0.42	7.126
4|5	4.6848	0.4683	10.003

Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's method ['lmerModLmerTest']

Formula: MCI ~ classification * target_IPA + (1 | word) + (1 + target_IPA | subject)
Data: fuldat

AIC	BIC	logLik	deviance	df.resid
1758.1	1977.9	-835	1670.1	1048

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.5116	-0.6986	-0.1117	0.596	6.4042

Random effects:

Groups	Name	Variance	Std.Dev.	Corr					
subject	(Intercept)	0.03746	0.19355						
target_IPAw		0.076236	0.27611	-0.52					
target_IPAt		0.017909	0.13382	-0.1	0.79				
target_IPAk		0.043052	0.20749	-0.03	0.79	0.89			
target_IPAj		0.007515	0.08669	0.19	-0.48	-0.3	-0.06		
target_IPAl		0.114693	0.33866	-0.63	0.93	0.66	0.77	-0.15	
target_IPAɹ		0.081878	0.28614	-0.33	0.71	0.55	0.33	-0.95	0.42
word	(Intercept)	0.00152	0.03898						
Residual		0.244766	0.49474						

Number of obs: 1092, groups: subject, 24; word, 16

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(>|t|)	
(Intercept)	**2.44347**	**0.06484**	**24.31562**	**37.686**	**< 0.0000000000000002**	*******
classificationSSD	0.06039	0.11615	23.28148	0.52	0.608031	
target_IPAw	**0.23011**	**0.10241**	**27.64664**	**2.247**	**0.032823**	*****
target_IPAt	-0.05244	0.074	17.93219	-0.709	0.487599	
target_IPAk	**0.30624**	**0.08123**	**28.21832**	**3.77**	**0.000768**	*******
target_IPAj	0.06587	0.09602	106.96647	0.686	0.494198	
target_IPAl	0.15931	0.11149	28.09393	1.429	0.164029	
target_IPAɹ	**0.42599**	**0.09405**	**27.85165**	**4.529**	**0.000101**	*******
classificationSSD:target_IPAw	0.02157	0.18179	30.89465	0.119	0.906316	
classificationSSD:target_IPAt	-0.09469	0.12901	22.65759	-0.734	0.470504	
classificationSSD:target_IPAk	-0.27862	0.14617	27.62854	-1.906	0.067091	.
classificationSSD:target_IPAj	0.04403	0.18953	190.20549	0.232	0.816547	
classificationSSD:target_IPAl	-0.04975	0.2008	24.61624	-0.248	0.806366	
classificationSSD:target_IPAɹ	-0.16532	0.16622	25.79062	-0.995	0.329188	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Cumulative Link Mixed Model fitted with the Laplace approximation

formula: NINFL ~ classification * target_IPA + (1 | word) + (1 + target_IPA | subject)
data: fuldat

link	threshold	nobs	logLik	AIC	niter	max.grad	cond.H
logit	flexible	1092	-1369.84	2831.68	7910(31550)	2.01E-03	8.40E+02

Random effects:

Groups	Name	Variance	Std.Dev.	Corr					
subject	(Intercept)	1.8087	1.3449						
target_IPAw		1.1964	1.0938	-0.522					
target_IPAt		0.7795	0.8829	-0.366	0.363				
target_IPAk		0.9444	0.9718	-0.633	0.775	0.52			
target_IPAj		0.8513	0.9227	-0.145	0.539	-0.47	0.474		
target_IPAl		1.7554	1.3249	-0.834	0.55	0.818	0.716	-0.17	
target_IPAɹ		1.9543	1.398	-0.779	0.528	0.254	0.229	-0.035	0.623
word	(Intercept)	0	0						

Number of groups: subject 24, word 16

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
classificationSSD	0.6264	0.69121	0.906	0.3648
target_IPAw	0.05817	0.41006	0.142	0.8872
target_IPAt	0.201	0.33854	0.594	0.5527
target_IPAk	0.34176	0.34649	0.986	0.324
target_IPAj	-0.01392	0.45178	-0.031	0.9754
target_IPAl	**0.98233**	**0.43179**	**2.275**	**0.0229***
target_IPAɹ	**1.26569**	**0.42028**	**3.012**	**0.0026****
classificationSSD:target_IPAw	-0.21983	0.74962	-0.293	0.7693
classificationSSD:target_IPAt	-0.68676	0.61296	-1.12	0.2625
classificationSSD:target_IPAk	-0.8752	0.64001	-1.367	0.1715
classificationSSD:target_IPAj	0.43282	0.83869	0.516	0.6058
classificationSSD:target_IPAl	-1.04922	0.79027	-1.328	0.1843
classificationSSD:target_IPAɹ	**-1.60918**	**0.76509**	**-2.103**	**0.0354***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Threshold coefficients:

	Estimate	Std. Error	z value
1|2	-0.5254	0.376	-1.397
2|3	1.1667	0.3774	3.091
3|4	3.0968	0.3895	7.952
4|5	4.8102	0.433	11.109

Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's method ['lmerModLmerTest']

Formula: MCI ~ transcription_accuracy_binary * classification * target + (1 | word) + (1 + target_IPA | subject)
Data: fuldat_rl

AIC	BIC	logLik	deviance	df.resid
647.2	696.9	-310.6	621.2	326

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.0299	-0.677	-0.114	0.5619	5.0734

Random effects:

Groups	Name	Variance	Std.Dev.	Corr
subject	(Intercept)	0.04842	0.22	
target_IPAɹ		0.07321	0.2706	-0.49
word	(Intercept)	0	0	
Residual		0.32936	0.5739	

Number of obs: 339, groups: subject, 24; word, 6

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	**2.74181**	**0.18117**	**156.23608**	**15.134**	**<0.0000000000000002*** **
transcription_accuracy_binarycorrect	-0.1583	0.19124	258.82902	-0.828	0.409
classificationSSD	-0.05577	0.24718	67.64444	-0.226	0.822
targetr	0.02148	0.19992	121.5473	0.107	0.915
transcription_accuracy_binarycorrect:classificationSSD	-0.02722	0.29131	119.66471	-0.093	0.926
transcription_accuracy_binarycorrect:targetr	0.37888	0.23244	227.26371	1.63	0.104
classificationSSD:targetr	0.09385	0.28321	53.23579	0.331	0.742
transcription_accuracy_binarycorrect:classificationSSD:targetr	-0.30396	0.35788	161.70451	-0.849	0.397

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Cumulative Link Mixed Model fitted with the Laplace approximation

formula: as.factor(NINFL) ~ transcription_accuracy_binary * classification * target + (1 | word) + (1 + target_IPA | subject)
data: fuldat_rl

link	threshold	nobs	logLik	AIC	niter	max.grad	cond.H
logit	flexible	339	-458	946	1129(2342)	1.24E-04	6.50E+02

Random effects:

Groups	Name	Variance	Std.Dev.	Corr
subject	(Intercept)	0.4549	0.6745	
target_IPAɹ		0.191	0.437	-0.771
word	(Intercept)	0	0	

Number of groups: subject 24, word 6

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
transcription_accuracy_binarycorrect	-0.4297	0.6063	-0.709	0.47846	
classificationSSD	-0.427	0.8243	-0.518	0.60446	
targetr	-0.6167	0.5949	-1.037	0.29988	
transcription_accuracy_binarycorrect:classificationSSD	0.1743	0.9696	0.18	0.85733	
transcription_accuracy_binarycorrect:targetr	**1.8623**	**0.703**	**2.649**	**0.00807**	******
classificationSSD:targetr	-0.3104	0.8695	-0.357	0.72106	
transcription_accuracy_binarycorrect:classificationSSD:targetr	-0.6478	1.1422	-0.567	0.57062	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Threshold coefficients:

	Estimate	Std. Error	z value
1|2	-1.5979	0.579	-2.76
2|3	-0.2379	0.5754	-0.413
3|4	1.7137	0.587	2.919
4|5	3.3308	0.6449	5.165

Word tokens elicited (random order)

barn
beard
board
butter
chair
cheer
clear
dark
door
farm
floor
flower
fork
hammer
ladder
mother
nurse
race
raft
rake
raw
reach
read
red
ride
ring
rip
rob
robe
rock
rose
rude
rug
rules
run
scare
scarf
share
sir
star
stare
stir
sword
tear
turn
weird
worm
wrap
wrong
year

/ɹ/ stimulability tokens elicited (3x each)

mer
der
erg
ree
ray
rai
roo
row
ra
ear
air
ayer
or
ar
our

/ɹ/ sentences elicited (random order)

I burned my finger on a roasting pan.

It's careless to leave rusty nails on the carpet.

Rachel got rid of the old fur coat.

Rob's old hat made his ear feel sore.

Sarah ran to the market to buy milk.

Varied phoneme tokens elicited (3x each; random order)

cape
cat
coat
key
lake
lamb
rake
rat
ring
rope
tea
toe
wake
wing
yam

Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's method ['lmerModLmerTest']

Formula: zScoreF3_F2 ~ scale(MCI, center = TRUE, scale = FALSE) * pre_post + (1 | word) + (1 | subject)
Data: fuldat_stim

AIC	BIC	logLik	deviance	df.resid
8228	8266.4	-4107	8214	1771

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.3134	-0.5564	-0.0281	0.4601	3.629

Random effects:

Groups	Name	Variance	Std.Dev.
subject	(Intercept)	9.01207	3.002
word	(Intercept)	0.05919	0.2433
Residual		5.52497	2.3505

Number of obs: 1778, groups: subject, 25; word, 15

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	**8.1713**	**0.613**	**26.6081**	**13.329**	**0.000000000000273*** **
scale(MCI, center = TRUE, scale = FALSE)	**0.5804**	**0.1536**	**1759.9559**	**3.779**	**0.000163*** **
pre_postPost	**-6.2336**	**0.129**	**1752.5945**	**-48.335**	**<0.0000000000000002*** **
scale(MCI, center = TRUE, scale = FALSE):pre_postPost	**-1.0136**	**0.176**	**1764.5566**	**-5.758**	**0.000000010004358*** **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's method ['lmerModLmerTest']

Formula: zScoreF3_F2 ~ NINFL * pre_post + (1 | word) + (1 | subject)
Data: fuldat_stim

AIC	BIC	logLik	deviance	df.resid
8265	8303.4	-4125.5	8251	1771

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.1859	-0.5504	-0.0427	0.4524	3.5559

Random effects:

Groups	Name	Variance	Std.Dev.
subject	(Intercept)	9.35762	3.059
word	(Intercept)	0.04464	0.2113
Residual		5.64684	2.3763

Number of obs: 1778, groups: subject, 25; word, 15

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	**7.98209**	**0.64712**	**30.61056**	**12.335**	**0.000000000000207*** **
NINFL	-0.03015	0.08681	1756.74869	-0.347	0.728
pre_postPost	**-6.26505**	**0.37652**	**1754.72403**	**-16.639**	**<0.0000000000000002*** **
NINFL:pre_postPost	0.06391	0.1398	1757.18305	0.457	0.648

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's method ['lmerModLmerTest']

Formula: MCI ~ scale(stereognosis, center = TRUE, scale = FALSE) * pre_post +
pre_post * scale(acuity_pre, center = TRUE, scale = FALSE) + (1 | word) + (1 | subject)
Data: ss_subset_stim

AIC	BIC	logLik	deviance	df.resid
3802.2	3851.6	-1892.1	3784.2	1769

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.9177	-0.6347	-0.0713	0.6113	4.1353

Random effects:

Groups	Name	Variance	Std.Dev.
subject	(Intercept)	0.430225	0.65592
word	(Intercept)	0.002723	0.05218
Residual		0.462042	0.67974

Number of obs: 1778, groups: subject, 25; word, 15

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	**2.75253**	**0.13551**	**26.59656**	**20.312**	**<0.0000000000000002*** **
scale(stereognosis, center = TRUE, scale = FALSE)	-0.06844	0.10561	26.56409	-0.648	0.523
pre_postPost	**0.43919**	**0.03512**	**1751.67226**	**12.505**	**<0.0000000000000002*** **
scale(acuity_pre, center = TRUE, scale = FALSE)	0.03655	0.0676	25.66053	0.541	0.593
scale(stereognosis, center = TRUE, scale = FALSE):pre_postPost	**-0.17358**	**0.02915**	**1752.07427**	**-5.955**	**0.0000000031291*** **
pre_postPost:scale(acuity_pre, center = TRUE, scale = FALSE)	**-0.11432**	**0.0167**	**1748.06807**	**-6.847**	**0.0000000000104*** **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Cumulative Link Mixed Model fitted with the Laplace approximation

formula: as.factor(NINFL) ~ scale(stereognosis, center = TRUE, scale = FALSE) *
pre_post + pre_post * scale(acuity_pre, center = TRUE, scale = FALSE) + (1 | word) + (1 | subject)
data: ss_subset_stim

link	threshold	nobs	logLik	AIC	niter	max.grad	cond.H
logit	flexible	1778	-2203.45	4428.89	1088(3267)	3.47E-03	6.40E+01

Random effects:

Groups	Name	Variance	Std.Dev.
subject	(Intercept)	0.3286	0.5732
word	(Intercept)	0.02036	0.1427

Number of groups: subject 25, word 15

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
scale(stereognosis, center = TRUE, scale = FALSE)	0.09752	0.11269	0.865	0.387
pre_postPost	**1.76966**	**0.10562**	**16.755**	**<0.0000000000000002*** **
scale(acuity_pre, center = TRUE, scale = FALSE)		-0.02353	0.0678	-0.347
scale(stereognosis, center = TRUE, scale = FALSE):pre_postPost	-0.01086	0.08072	-0.135	0.893
pre_postPost:scale(acuity_pre, center = TRUE, scale = FALSE)	-0.05403	0.04623	-1.169	0.242

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Threshold coefficients:

	Estimate	Std. Error	z value
1|2	-0.8638	0.1479	-5.841
2|3	0.9014	0.1499	6.012
3|4	3.2403	0.1659	19.531
4|5	5.2355	0.223	23.475