INDIVIDUAL DIFFERENCES IN THE ACQUISITION OF THE /t/ - /k/ CONTRAST: A STUDY OF ADULTS' PERCEPTION OF CHILDREN'S SPEECH

A THESIS
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Sara Rose Bernstein

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF ARTS

Benjamin Munson, Ph.D.

May 2015

Acknowledgements

I would like to acknowledge the Learning to Talk grant supported by the National Institute for Deafness and Other Communicative Disorders (NIDCD 02932) and the National Science Foundation, for providing funding to carry out this research endeavor. Greatest thanks are owed to my advisor, Dr. Benjamin Munson for his unwavering support and guidance throughout this project, without whom this thesis would not have been written.

This project was guided by the input of Dr. Mary Beckman at the Ohio State University and Dr. Jan Edwards at the University of Wisconsin–Madison. I give special notice to the Learning to Talk lab manager at the University of Minnesota, Dr. Maria Swora for truly investing herself in the project.

Thanks are due to those involved in developing stimulus preparation scripts. Dr. Mary Beckman was a source of wisdom throughout this project, providing particular support in improving the burst tagging process, developing stimulus preparation scripts, and guiding me through the initial stages of data analysis. Dr. Franzo Law II tailored the burst tagging script and responded to all of my tagging questions in an exceedingly timely and calming manner. Pat Reidy contributed significantly to the stimulus preparation scripts. Thank you, Dr. Mary Beckman and Pat Reidy for developing the robustness-of-contrast analysis code.

Data collection and analysis for this project were undoubtedly a team effort. I owe great thanks to each of the individuals listed below, for their skill, talent, attention to detail, and dedication to the project. A wonderful team of examiners collected data (in alphabetical order): Jamie Anderson, Natasha Arora, Tatty Bartholomew, Ruby Braxton, Eileen Brister, Nicole Bruenig, Jamie Byrne, Marcy Campbell, Kareem Darwiche, Cara Donohue, Dana Duncan, Tyler Ellis, Kerri Engel, Michelle Erskine, Colette Felion, Megan Flood, Allison Holt, Courtney Huerth, Allison Johnson, Kelly Jorgensen, Isla Katz, Kayla Kristensen, Franzo Law II, Daria Lawrence, Annie Loof, Tristan Mahr, Sarah McGowan, Morgan Meredith, Michelle Minter, Yakira Moore, Amy Muzynoski, Hannele Nicholson, Jill Pettit, Mandi Proue, Danielle Revai, Erica Richmond, Sarah Schellinger, Alissa Schneeberg, Bianca Schroeder, Janet Schwartz, Malia Silvert, Maria Swora, Tatiana Thonesavanh, Nancy Wermuth, Colleen Woyach.

The data analysis teams segmented the children's recordings: Jamie Anderson, Cara Donohue, Tyler Ellis, Megan Flood, Rose Janecke, Amy Muzynoski, Hannele Nicholson, Bianca Schroeder, Janet Schwartz, Kristi Warndahl, Haley Webb, and checked the segmented text grids: Rose Janecke, Allison Johnson, Mia Kim, Hannele Nicholson, Haley Webb. Thanks to my burst-tagging partner, Allison Johnson for taking this adventure with me. The perception testing team, who collected data from adult listeners, were a tremendous support: Olivia Cox, Emma Hage, Lauren Kosky, and Mara Logerquist.

Finally, I am grateful to all of the families involved in the Learning to Talk project, as well as the listener participants for this perception study, for being so generous of their time.

Abstract

The presence of subtle but meaningful within-category sound differences has been documented in acoustic and articulatory analyses of children's speech. This study explored visual analog scaling (VAS) to measure speech perception. Productions of word-initial /t/ and /k/ were recorded from a diverse group of 63 children aged 28 to 39 months. Adult naïve listeners rated productions on a VAS. Measures of children's vocabulary, speech perception, executive function, home language environment, and maternal education level were collected. Robustness of the /t/-/k/ contrast was derived from adult VAS ratings for each talker. Speech accuracy, based on phonetic transcriptions was calculated. Listeners differentiated transcription categories, including intermediate categories, using the VAS. Listeners had variable levels of intra-rater reliability, and set effects were present. Transcription accuracy and robustness of contrast were closely related, but robustness of contrast highlighted differences between children with high accuracy. Vocabulary measures predicted both robustness of contrast and transcribed accuracy.

Table of Contents

List of Tables

List of Figures

1 Introduction

Phonological knowledge is multifaceted. It involves knowledge of the way that sounds are produced, how they are perceived, and ways that variations in sound are used to convey meaning in a language (Munson, Edwards & Beckman, 2005). Studies of each of these areas can yield valuable information about phonological development and more general language development. The current study focused on just one of these facets, sound production. This choice was made for a variety of reasons. Both standardized and informal measures of phoneme production accuracy are thought to have especially high ecological validity, as they are seen as a measures of what others observe the child to do when speaking. Such measures are widely understood to represent a child's speech and language development in both clinical and general settings (Bleile, 2002; Khan, 2002; Tyler, & Tolbert, 2002). Production accuracy can be measured through a variety of methods, including listener perception, articulatory analysis, and acoustic analysis. Using these techniques in tandem provides the opportunity for cross-validation of acoustic, articulatory, and perceptual measures.

The current study investigated the development of the production of /t/ and /k/ in children aged 28 to 39 months. For many children, adult-like production of these sounds emerges but is not mastered during this interval (Smit, Hand, Freilinger, Bernthal, & Bird, 1990). The specific sound contrast was chosen because it is commonly produced in error in younger typically developing children and in older children with speech sound disorders. English speaking children may tend to produce errors on the /k/ phoneme that resemble correct production of /t/ (Beckman, Munson, & Edwards, 2014; Stoel-Gammon,

1991).

Both /t/ and /k/ are produced by stopping the outflow of air from the vocal tract with the a closure of the tongue at the alveolar ridge (/t/) or the soft palate (/k/), and then releasing the air in a burst. There are many possible places of articulation for the tongue between the anterior (alveolar) and posterior (velar) sites. This range in place of articulation correlates to a range in possible acoustic outputs for these attempts to produce /t/ or /k/ (Edwards, Gibbon, & Fourakis, 1997; Forrest, Weismer, Hodge, & Dinnsen, 1990). Indeed, studies using acoustic analysis or direct articulatory measurements have found that some children produce sounds that are intermediate between /t/ and /k/. These within-category sound differences have been referred to as *covert contrasts*. Children who produce covert contrasts may use different articulatory gestures to produce /t/ and /k/, but the acoustic outputs that they produce are denoted by the same symbol in phonetic transcription (Forrest, Weismer, Hodge, Dinnsen, & Elbert, 1990; Gierut & Dinnsen, 1986; Macken & Barton, 1980).

Early evidence of covert contrasts was documented by Macken and Barton (1980). These researchers examined four children's (aged one year, four months to one year, seven months at onset of study) productions of word-initial stop consonants (/p, b, t, d, k, g/). Recordings were made every two weeks over an eight month period. All productions were transcribed, and four sets of productions from each child were analyzed acoustically. Voice onset time (VOT) of the word-initial consonant was calculated. VOT refers to the duration of the interval between the release of a stop consonant closure and the onset of vocal-fold vibration in the following vowel. In English, VOT is the primary

cue to the voicing contrast, with so-called 'voiced' stops being produced with a very short-lag VOT (i.e., voicing begins simultaneous with or shortly after the release of the consonant) and voiceless sounds being produced with a long-lag VOT (i.e., there is a substantial interval between the release of the stop consonant closure and the onset of voicing in the following sound). The timing of the release of a stop consonant and the onset of voicing in the following vowel is relatively unconstrained biomechanically, meaning that there are many possible values of VOT for a given consonant. The VOTs of children's productions were categorized by Macken and Barton into one of three categories. Category 1 data included productions for which there was no differentiation in VOT between the productions of voiced and voiceless targets. Category 2 included productions where target voiced and target voiceless sounds were produced with different VOTs, but the contrast was not yet adult-like. Specifically, these productions fell within the voiced category (for the target voiced sounds) or around the adult perceptual boundaries of the VOT contrast (for the voiceless targets). Perceptual boundaries are the point on a continuum where an adult changes the label they give to a sound in a two alternative forced choice task. These sounds were transcribed as voiced, despite the fact that the voiced and voiceless targets were produced differently. Category 3 included adult-like voicing contrasts. The Category 2 data demonstrated that phonetic transcriptions might mischaracterize children's productions. Sound contrasts do not change from being absent to present in a discontinuous manner. Rather, acoustic markers emerge continuously over time.

More recent studies have found evidence of covert contrasts for other sounds. An

acoustic analysis by Forrest, Elbert, Weismer, and Dinnsen (1994) compared /t/ and /k/ productions of three groups of children: typically developing children who had mastered the /t/-/k/ contrast, children with a phonological disorder who had mastered the /t/-/k/ contrast fully (as assessed by phonetic transcriptions of words with /t/ and /k/ in a variety of word positions), and children with a phonological disorder who only produced a correct /t/-/k/ contrast at the beginnings of words. The researchers found that /t/ and /k/ productions of the former two groups (the typically developing children and the children with phonological disorder who had mastered the /t/-/k/ contrast in all word positions) were acoustically distinct, and were produced similarly by both groups. However, the acoustic characteristics of /t/ and /k/ productions from the latter group were less distinct from one another, even though they had been transcribed as correct. This shows that there are a wide range of acoustic outputs possible within the perceptual boundaries of /t/ and /k/. Additionally, the degree or robustness of contrast between productions is meaningful in determining the child's level knowledge about the sounds.

Covert contrasts can also be documented with direct articulatory measures. Gibbon (1990) employed electropalatography, a tool that measures and displays the contact of the tongue to the top of the mouth (palate). Gibbon investigated tongue-palate contact during /d/ and /g/ production. The /d/ and /g/ are characterized by the same articulatory posturing as the /t/ and /k/, respectively. The sound pairs differ in VOT, with /t/ and /k/ having long-lag VOTs and /d/ and /g/ having short-lag VOTs. Gibbon studied two children, whose productions of target /d/ and /g/ were transcribed to be identical, as well as one adult speaker with typical productions of these sounds. The

electropalatography data showed that the children clearly and consistently differentiated between target /d/ and target /g/, although in different ways from each other and from the adult speaker. These results provide support for the existence of covert contrasts in the production of stop consonant place. Hence, they provide further evidence of the gradient nature of phoneme acquisition, and the shortcomings of the forced-choice system of phonetic transcription.

Following documentation of covert contrasts, researchers have studied the clinical significance of these subtle, within-category acoustic differences. Gierut and Dinnsen (1986) transcribed and acoustically analyzed the sound productions of two children who appeared, prior to analysis, to be producing the same sound error pattern. When analyzed, the authors found that one child was truly not marking sound contrasts, while the other child was marking contrasts in consistent yet subtle ways. They noted the importance of an alternative method to phonetic transcription in understanding the sound production knowledge of these two children. Further, though the phonetic transcriptions of these two children suggested that they had similar treatment needs, the acoustic analysis suggests that they might benefit from different therapeutic goals and teaching approaches. Tyler, Figurski, Langsdale (1993) investigated the relationship between covert contrasts and progress in speech therapy. The authors noted one participant who, prior to treatment, produced a "significant acoustic distinction" between velar /k, g/ and alveolar /t, d/ target sounds that was "largely imperceptible" (p. 747). Even though the child was producing acoustic differentiation, phonetic transcription suggested that he was producing the targets incorrectly and indistinctly. This child made some of the fastest and most

significant gains over the treatment period compared to the six other participants. This prompted the authors to conclude that the presence of some acoustic differentiation may facilitate faster learning and generalization of a sound target. MacLeod and Glaspey (2014) found that children with speech sound disorders gradually progressed toward producing more acoustically velar-like sounds over the course of speech therapy. As the children improved their ability to produce acoustically velar-like sounds, they required less cueing to produce velar stops. Acoustic analysis and required cueing level captured the gradual process of sound acquisition that phonetic transcriptions did not encode. Taken together, these studies illustrate that the presence of covert contrasts can play a significant role in goal setting, treatment approach, and progress in therapy.

Despite the mounting evidence for gradient differences in speech sound production, traditional clinical and research methods for measuring sound accuracy and development have relied on the binary correct or incorrect measures based on phonetic transcriptions (Gardner, 1997). Indeed, a great deal of the scientific knowledge about children's speech sound production, including age of acquisition norms, has been built upon phonetic transcriptions (Smit et al., 1990; Macken & Barton, 1980; Gierut & Dinnsen, 1986). Phonetic transcriptions cannot document the possible range in output shown in articulatory and acoustic studies of children's speech (Forrest et al., 1990; Forrest et al., 1994; Gierut & Dinnsen, 1986; Li, 2012). Some studies employ intermediate phonetic transcription categories (e.g., Munson, Edwards, Schellinger, Beckman, & Meyer, 2010, based on the suggestion by Stoel-Gammon, 2001). A transcription of [s:ʃ] indicates a sound perceived closer to /s/, but with some /ʃ/-like

("sh"-like) qualities. Even with the use of intermediate categories, transcriptions do not capture all of the within-category detail that might be relevant for understanding phonological acquisition and disorders. To study gradual acquisition, a tool must be able to capture fine-grained differences in speech sound production. Gibbon (1990) remarked that transcription is an "oversimplification or even misrepresentation" (p. 338) of a child's sound production knowledge. Moreover, transcription is subject to sometimes idiosyncratic individual differences related to individuals' unique perceptual abilities and linguistic histories (Ladd, 2011).

On a psychometric level, both trained and untrained listeners fail to achieve acceptable levels of intra- and inter-rater reliability using a two-alternative forced choice paradigm (Mayo, Gibbon, & Clark, 2013). Mayo et al. (2013) presented consonant-vowel-consonant sequences ("a go" and "a doe") to trained and untrained listeners. The stimuli were synthetic speech, with the transitions into and out of the consonants manipulated to produce both clear /d/ and /g/ sounds and intermediates. Both groups of listeners had poor intra-rater reliability. The authors remarked "listeners had high perceptual sensitivity to within-category detail but difficulty pairing that sensitivity with the limited number of categories provided for them" (p. 786).

Because speech contains within-category acoustic differences, some talkers may produce sounds in a contrast more similarly than other talkers. It is meaningful to characterize a talker's degree of difference, or robustness of contrast, between two sounds. A talker who consistently produces /t/ and /k/ distinctly is said to have a robust contrast for this pair. Robustness of contrast has been applied to acoustic measures from

the /s/ and /ʃ/ sounds (Holliday, Reidy, Beckman, & Edwards, 2014; Perkell, Matthies, Tiede, Lane, Zandipour, Marrone, Stockman, & Guenther, 2004; Romeo, Hazan, & Pettinato, 2013). Perkell et al. (2004) used separation of spectral mean, the average of frequency components in a sound, to characterize robustness of the /s/ - /ʃ/ contrast. Romeo et al. (2013) proposed three measures of robustness of contrast: within-category dispersion (degree of spread around the spectral mean), between-category distance (difference in spectral mean values between the two sounds), and discriminability, *d*(a), (between category distance divided by the square root of mean variances). Holliday et al. (2014) introduced *percent correctly predicted*, which inputs spectral mean into a model of the /s/-/ʃ/ characteristics. The model predicts whether the production was transcribed as [s] or [ʃ]. The proportion that matches transcription category for each talker yields percent correctly predicted. This measure describes category overlap, like between-category distance, but without the influence of distance in separation.

Acoustic analysis has played a significant role in contributing to the speech sound literature (Forrest et al., 1990; Forrest et al., 1994; Gierut & Dinnsen, 1986). At this time, however, acoustic measures remain imperfect and largely impractical for clinical use. Acoustic measures do not consistently correspond to articulatory gestures (Marin, Pouplier & Harrington, 2010). Most clinical settings lack the equipment and quiet environments to gather high quality recordings, and the analysis process can be time consuming. Similarly, the instrumentation to conduct electropalatography remains prohibitively costly for the majority of therapy settings. The limitations of acoustic and articulatory analysis, and perceptually based phonetic transcriptions establish a need for a

clinically viable, acoustically valid, and perceptually reliable tool to capture within-category variation in children's productions.

One solution to this problem is to use continuous rating scales. A growing number of studies have investigated listener perception of covert contrasts using visual analog scales (Beckman, Munson, & Edwards, 2014; Julien & Munson, 2012; Munson et al., 2010; Munson, Johnson, & Edwards, 2012; Strömbergsson, 2014). Visual analog scaling (VAS) refers to any method in which a sensory percept is converted into a visual analog of that percept. One widely used VAS is for the assessment of pain in emergency medical settings. In these scales, pain is represented visually by simple facial expressions. People select the facial expression that corresponds to the level of pain they are experiencing (DeLoach, Higgins, Caplan, & Stiff, 1998)

In the previous research most relevant to the current, the VAS was a double-headed arrow with a hash mark in the center of the line. At each endpoint of the line, the text "the 'X' sound" was written. Listeners heard a sound and clicked along the line where they perceived the sound to fall. The click location was taken as the rating. The text associated with 'X' varied across studies: in Julien and Munson (2012), it was 's' at one end and 'sh' at the other. In Munson et al. (2010), it was 's' and 'th'. The double-headed arrow with a mark at its center was meant to invoke the number line: a neutral midpoint (as in the case of the number zero), and continuous variation (as suggested by the arrowhead) away from the neutral midpoint, toward either one sound or the other.

Studies of adults' perception of children's productions of the voiceless lingual fricatives /θ/, /s/, and /ʃ/ using VAS have argued that ratings using this technique are a

viable proxy for acoustic analysis. Several pieces of evidence support this claim. First, both trained and untrained listeners provide a variety of VAS ratings for sounds that are transcribed with the identical phonetic symbol, and these correlate with acoustic characteristics of the sounds being rated (Julien & Munson, 2012; Munson, Johnson, & Edwards, 2012). When compared to other perceptual tools, such as reaction time to a forced choice and direct magnitude estimates of category goodness, VAS ratings have superior intra-rater reliability, and have a stronger correlation with sounds' acoustic characteristics (Munson & Urberg Carlson, 2015). While most of the studies cited above examined the perception of children's fricatives, a small number of recent studies have suggested that VAS is also a useful tool for measuring acquisition of the /t/ - /k/ contrast. Studies using acoustic analysis have shown that this contrast is acquired gradually (Forrest et al., 1990; Forrest et al., 1994). Munson et al. (2012) showed that sounds transcribed as intermediate between /t/ and /k/ are given VAS ratings that are intermediate between those for /t/ and those for /k/. Strömbergsson (2014) extended this finding to Swedish, and showed that VAS ratings correlate with acoustic characteristics of the sounds being rated. Beckman, Munson, and Edwards (2014) showed that Japanese- and English-speaking adults' VAS ratings of Japanese- and English-acquiring children's /t/ and /k/ productions differed in ways that are predicted by cross-linguistic differences in the production of the /t/-/k/ contrast.

In sum, VAS is a potentially powerful tool because of its simplicity and utility in describing, through the ecologically valid metric of listener perception, a child's production accuracy and progress in speech therapy. VAS can also be used to provide a

robustness of contrast measure without the need for acoustic analysis. Robustness of contrast can be defined in terms of separation of ratings along the VAS between contrasting sounds, and capture the variability in ratings for each sound. Further investigation of its use is necessary to determine whether VAS is a clinically viable method for capturing subtle acoustic differences in the /t/-/k/ contrast.

1.1 Aims of this study

The current study had two general aims. The first aim was to develop and validate a VAS to measure children's productions of words that began with a /t/ or a /k/ target. Like previous scales, this should elicit a continuous response, thereby allowing the tool to measure the gradual acquisition of this contrast. Given the previous findings by Beckman et al. (2014), Munson et al. (2012), and Strömbergsson (2014), it was hypothesized that untrained adult listeners will utilize the full range of a visual analog scale when presented with children's productions. We predicted that these ratings would utilize the entire range of the VAS, that they would have a high degree of intra-rater reliability, and that they would differentiate among different transcription categories, including among both endpoint transcriptions and intermediate ones. It was hypothesized that listener's responses would vary significantly based on transcription category, with greater variance in the responses described by the intermediate transcription categories.

The second aim of this study was to derive measures of how robustly children's /t/ and /k/ productions differed based on listeners' ratings, and to examine predictors of child-by-child differences in the VAS-derived measures of robustness of /t/-/k/ contrast. A large set of measures, beyond the speech production samples used as stimuli in the

VAS studies, was collected from the children as part of a larger longitudinal study. Predictor variables included ones that were related to input (home language environment, maternal education, dialect of English spoken at home, status as a late talker) or output related (vocabulary, executive functions, speech sound discrimination). It was hypothesized that both input and output related factors would play a significant role in predicting children's robustness of contrast for /t/ and /k/. Because the productions that were used as stimuli in this study were phonetically transcribed, this study also provided the opportunity to examine whether the VAS-derived measures of robustness of contrast were predicted by a different set of measures than those that predicted accuracy as determined by phonetic transcription. A finding that different factors predicted the two measures of /t/-/k/ production would suggest that they index different underlying skills.

2 Methods

This section is divided into three subsections. The first describes the *child talkers* whose productions were used as stimuli in the VAS perception study. It includes a description of the characteristics of the talker participants, the predictor variables identified in this study, and the methods for collection speech samples. The second section describes the procedures for stimulus preparation. The third section describes the *adult listeners* and the procedure for the perception study.

2.1 Child talkers

This section presents characteristics of the talkers. Child-related output and input predictors of speech production are discussed. A table of correlations to characterize the

set of predictor variables is shown. Finally, the procedure for speech sample collection is described.

2.1.1 Talker participants

The stimuli for this perception study were produced by 63 children, 28 to 39 months old (Figure 1). Children were recorded at both the University of Minnesota in Minneapolis and the University of Wisconsin in Madison. All children passed a hearing screening of 1000, 2000 and 4000 Hz tones presented at 25 dB HL. The children were recruited via advertisements in local newspapers, connections with community organizers, and fliers posted around the community.

All children included in this thesis participated as part of a larger longitudinal study on development of phonological knowledge and vocabulary (www.learningtotalk.org). Testing was completed over two or three visits, of one to two hours each. The children were all from monolingual, English-speaking households per

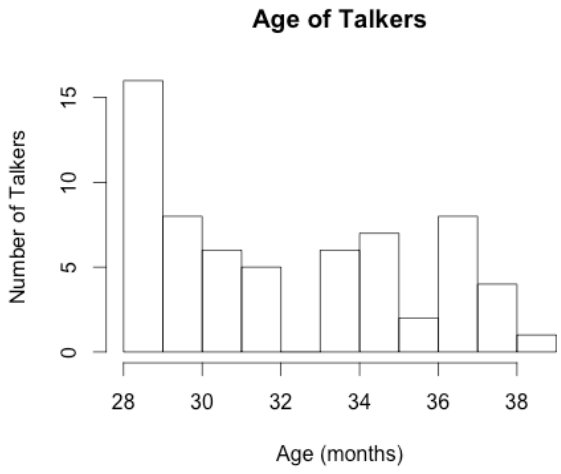

Figure 1: Distribution of talker ages

caregiver report. This study included children from both Mainstream American English (MAE) and African American English (AAE) dialect home language environments. Dialect was determined during a pre-visit phone interview and confirmed at the first testing session. Morphological and phonological elements of AAE were considered. Late talker status was defined as receptive vocabulary and prelinguistic skills within normal limits, with expressive vocabulary below normal limits for a child's age, with no other speech, language, hearing, or developmental diagnoses. Talker participants represented a range of maternal education levels. A table of child characteristics is shown below (Table 1 through Table 3). Due to the large number of talkers and speech tokens used as stimuli in this study, talkers were assigned to one of three different experiment versions (A, B, C). This ensured that no one listener would participate in an overly long experiment. Talker assignment was balanced by age, sex, maternal education, late talker status, and dialect (Table 4). In order to identify predictors of speech production ability, a variety of measures was collected to describe child-level differences in language and related areas. This project categorizes these talker-related variables as either output variables (i.e., measures of individual children's performance), or input variables (i.e., home language environment).

Table 1: Child talkers in experiment version A

Talker ID	Age (months)	Sex	Maternal Education	Late Talker	Dialect
001L	28	F	College degree	No	MAE
010L	32	M	Graduate degree	No	MAE
013L	32	M	GED	No	AAE
014L	39	M	Graduate degree	No	MAE
025L	37	F	High School diploma	No	AAE
033L	35	F	College degree	No	MAE
046L	35	F	Some college	No	AAE
049L	38	M	College degree	Yes	MAE
051L	29	F	Some college	No	MAE
058L	36	F	Some college	No	MAE
086L	35	M	Some college	No	MAE
087L	30	M	College degree	No	MAE
108L	29	M	Graduate degree	Yes	MAE
131L	32	M	Some college	No	MAE
133L	35	M	Trade school OR Associate's/ Technical OR Some college	No	MAE
604L	30	F	Graduate degree	No	MAE
607L	36	M	College degree	No	MAE
613L	34	F	Graduate degree	No	MAE
620L	28	F	College degree	No	MAE
660L	28	F	Graduate degree	No	MAE
671L	31	M	Some college	Yes	AAE
675L	31	F	College degree	No	MAE

Table 2: Child talkers in experiment version B

Talker ID	Age (months)	Sex	Maternal Education	Late Talker	Dialect
006L	28	F	College degree	No	MAE
012L	30	M	Graduate degree	No	MAE
017L	28	M	GED	No	AAE
034L	38	F	Trade school OR Associate's/ Technical OR Some college	No	MAE
035L	32	F	Some college	No	AAE
039L	37	M	College degree	No	MAE
051L	29	F	Some college	No	MAE
066L	38	F	High School diploma	No	AAE
088L	30	M	Trade school OR Associate's/ Technical OR Some college	No	MAE
089L	29	M	Graduate degree	Yes	MAE
092L	37	M	High School diploma	No	MAE
107L	34	F	Some college	No	MAE
123L	28	M	Graduate degree	Yes	MAE
600L	37	M	Graduate degree	No	MAE
610L	31	F	Graduate degree	No	MAE
629L	30	M	Graduate degree	No	MAE
630L	28	F	Trade school	No	MAE
632L	37	F	Graduate degree	No	MAE
646L	34	M	GED	No	AAE
661L	28	F	Some college	No	MAE
673L	35	M	College degree	No	MAE
680L	31	M	Graduate degree	Yes	MAE

Table 3: Child talkers in experiment version C

Talker ID	Age (months)	Sex	Maternal Education	Late Talker	Dialect
022L	34	F	Graduate degree	No	MAE
024L	31	M	Trade school	No	AAE
036L	29	F	Some college	No	AAE
040L	37	F	High School diploma	No	MAE
051L	29	F	Some college	No	MAE
053L	35	M	Trade school OR Associate's/ Technical OR Some college	No	MAE
067L	37	F	High School diploma	No	AAE
071L	30	M	Graduate degree	No	MAE
076L	34	M	College degree	No	MAE
083L	30	F	Some college	No	MAE
093L	28	F	Some college	Yes	MAE
101L	38	M	College degree	No	MAE
110L	30	M	College degree	No	MAE
128L	31	F	College degree	Yes	MAE
602L	34	M	Graduate degree	No	MAE
603L	35	F	Graduate degree	No	MAE
612L	29	F	Technical/Associate’s degree	No	MAE
636L	29	F	Trade school OR Associate's/ Technical OR Some college	No	MAE
640L	37	F	High School diploma	No	MAE
655L	28	M	Graduate degree	No	MAE
681L	32	F	Graduate degree	Yes	MAE

Table 4: Child talker information by experiment version. Maternal education was converted to an ordinal variable for each child, averaged, and then converted back to a nominal value.

Experiment Version	Number of talkers	Average age (months)	Number of females	Average maternal education	Number of late talkers	Number of AAE speakers
A	22	32.7	11	Some college to College degree	3	4
B	22	32.2	10	Some college to College degree	3	4
C	21	32.2	13	Some college to College degree	3	3

2.1.2 Output predictor variables

Output measures were a series of standardized and non-standardized assessments. They included both experimenter-administered activities and questionnaires completed by the child's caregiver. These were measures of executive function, speech perception, and vocabulary knowledge (Table 5). Tests were administered by trained undergraduate and graduate students.

Table 5: Predictor output measures, shown with label used in correlation tables

Measure (Label used to refer to this measure in the Results section)	**Description**
Fruit Stroop (FruitStroop)	Attention and inhibition, tested
Behavioral Rating Inventory of Executive Functions global composite percentile (BREIFGlobal Percentile)	Overall measure of executive functions, caregiver questionnaire
Minimal Pair Discrimination (MinPairs)	Percentage of pictures correctly identified in auditory-based minimal pair discrimination task, field of 2 pictures, tested
Expressive Vocabulary Test Raw Score (EVT_Raw)	Raw score on standardized, norm-referenced assessment of expressive vocabulary, tested
Expressive Vocabulary Test Standard Score (EVT_Stnd)	Standard score on standardized, norm-referenced assessment of expressive vocabulary, tested
Expressive Vocabulary Test Growth Scale Value (EVT_GSV)	Growth scale value on standardized, norm-referenced assessment of expressive vocabulary, tested
Peabody Picture Vocabulary Test Raw Score (PPVT_Raw)	Raw score on standardized, norm-referenced assessment of receptive (understanding) vocabulary, tested
Peabody Picture Vocabulary Test Standard Score (PPVT_Stnd)	Standard score on standardized, norm-referenced assessment of receptive (understanding) vocabulary, tested
Peabody Picture Vocabulary Test Growth Scale Value (PPVT_GSV)	Growth scale value on standardized, norm-referenced assessment of receptive (understanding) vocabulary, tested
MacArthur-Bates Communication Development Inventory total number of words (CDI_Produce)	Number of words child produces across environments, caregiver questionnaire

Executive function measures were used to investigate whether children's ability to attend to relevant information and inhibit extraneous information plays a significant role in speech production abilities. The Fruit Stroop test was used, in which a child saw a small fruit (apple, orange, banana) inscribed in a larger different fruit (similar to

Archibald & Kerns, 1999). The child must attend to the small fruit while inhibiting the competing information of the larger fruit. The BRIEF is a caregiver-completed questionnaire that asks questions regarding the child's behavioral regulation and metacognition (Gioia, Espy, & Isquith, 2003).

Speech perception was assessed through a minimal pair picture discrimination task. In this activity the child heard one word over speakers and was then presented with two pictures, one of the spoken word, and one of a word that differed by one speech sound (e.g. "bear" played over speakers, pictures of "bear" and "pear" presented on screen). Participants responded via touch screen. Investigating speech perception is important because many speech sound production errors are rooted in phonological perception difficulties (Locke, 1980). Further, speech perception abilities provide insight into a child's phonological system. Because better accuracy at identifying the labeled word of two minimal pairs indicates greater phonological knowledge, it was hypothesized that children with higher scores would produce a more adult-like /t/-/k/ contrast.

Vocabulary size has been shown to be a strong predictor of some aspects of phonological knowledge in children (Edwards, Beckman & Munson, 2004; Stoel-Gammon, 1991). In order to explore this relationship more fully, multiple measures of vocabulary knowledge were included. Vocabulary was measured via administration of the Expressive Vocabulary Test – 2nd Edition (EVT-2, Williams, 2007), for vocabulary production, and the Peabody Picture Vocabulary Test – 4th Edition (PPVT-4, Dunn & Dunn, 2007), for vocabulary understanding. Tests were administered in accordance with standardized protocols. Additionally, the MacArthur Bates Communication Development

Inventory, a parent-completed questionnaire, was completed to identify the total number of words the child produces across environments (Fenson, Marchman, Thal, Dale, Reznick, & Bates, 2007). It was hypothesized that children with higher vocabulary scores on all measures would produce more robust contrasts.

2.1.3 Input predictor variables

Child-level input predictor measures were collected through surveys and Language Environment Analysis Pro (LENA Pro) recordings. Surveys were completed by the caregiver. Survey data were maternal education level, as an index for socioeconomic status (SES), status of late to start talking, and dialect. In the current study, maternal education level was interpreted as an ordinal variable (1=GED or high school diploma, 2=technical, trade school or Associate's degree, 3=some college, 4=college degree, 5=graduate degree). The home language measures were recorded with LENA digital language processors and accessed through LENA Pro software. Measures were adult words heard, conversational turns, and meaningful speech heard (Table 6).

Table 6: Predictor input measures, shown with label used in correlation tables

Measure	Description
Adult words heard per hour (WordsPerHour)	Number of words spoken by an adult near the child per hour, LENA measure
Conversational turns per hour (CTCPerHour)	Number of turns in a conversation a child takes per hour, LENA measure
Minutes of meaningful speech (Meaningful)	Minutes of meaningful speech a child is exposed to per hour, LENA measure
Maternal education level (MatEdOrdil)	Expressed as ordinal variable with levels: GED, High school diploma, Some college, trade/technical school or Associate's degree, College degree, Graduate degree
Late Talker	Child was late to start talking with no other speech, language, development, or hearing diagnoses, parent report and tested
Dialect	Dialect of English spoken at home: African American English, Mainstream American English

Walker, Greenwood, Hart, and Carta (1994) reported on the language and intelligence outcomes of school aged-children, for whom SES, intelligence, and home language environment were measured between seven and 36 months. The authors found that at seven to 36 months, children from higher-SES households were exposed to a greater variety of words than their peers from lower-SES households. Children's receptive and expressive language test scores, including receptive vocabulary, measured seven years later were strongly correlated to SES, language input, and intelligence measures from early in life. To examine the relationship between SES and the sound system, Nittrouer (1996) measured children's phonological knowledge with two sound manipulation tasks. Four groups were compared: mid-SES with no history of ear

infections, mid-SES with chronic ear infections, low-SES with no history of ear infections, and low-SES with chronic ear infections. The children from low-SES backgrounds performed similarly on tasks to children from mid-SES backgrounds with frequent ear infections. All groups performed more poorly than the children from mid-SES backgrounds with no history of ear infections. It was hypothesized that children from higher-SES households in this study would produce more robust contrasts.

The LENA Pro system was used to collect data on children's language exposure on a typical day (Gilkerson & Richards, 2009). Each child wore a digital language processor recording device for one full day. The recordings were then processed with LENA Pro software, yielding daily total and hourly information on adult word count (number of words spoken by an adult in proximity to the recorder), conversational turn count (number of back and forth conversational exchanges between the child and adult), and percent of meaningful speech. It was hypothesized that the children who were exposed to richer linguistic environments would produce more robust sound contrasts.

2.1.4 Correlations between predictor variables

Because many of the predictor variables describe similar constructs (e.g. both the EVT scores and the CDI measure expressive vocabulary), it was expected that several of the input variables would be highly correlated. Descriptive correlations are presented below to characterize the set of independent variables. Because many of these variables are also correlated with age, both full correlations and partial correlations are presented. Partial correlations, in which the variables are residualized for age (i.e. the effect of age is removed) are presented above the diagonal, while full correlations are below the diagonal

(Table 7).

2.1.5 Speech production data collection

The stimuli for this perception study were recorded during a picture-based auditory word repetition activity. This task was administered via a computer running E-Prime software. Auditory prompts were presented from Klipsch BT77 speakers, normalized to 70 dB, in a sound-treated booth. Speech recordings were collected with an Audio Technica (AT 4040) cardioid capacitor microphone and a Marantz Professional solid state recorder (PMD671).

Speech production data were collected by trained undergraduate and graduate students. A visual reinforcer of an animal climbing a ladder was employed to increase motivation, in addition to praise, encouragement, and stickers. A total of 99 test trials were presented during this task, and were selected to be familiar to young children. For this study, 17 initial /t/ and /k/ target words were interspersed with the other targets for studies on speech sound development. Stimuli were presented in a randomly shuffled order for each participant. There were 17 target words (eight alveolar initial, nine velar initial) with two productions elicited for each target. The targets were selected to include high front, high back, and low back vowel contexts.

/t/: tummy, table, toast, tooth, tongue, tape, teddy bear, tickle

/k/: kitty, kitchen, candy, coat, car, cake, cup, cat, cookie

Table 7: Coefficients for correlations between independent measures. Partial correlations, without the effect of age on each variable, are shown above the diagonal. Full correlations, including the effect of age on each variable, are shown below the diagonal.
*** = p < 0.05 ** = p < 0.01**

Control Variables		Age	Fruit Stroop	BRIEF Global Percentile	Min Pairs	EVT Raw	EVT Stnd	EVT GSV	PPVT Raw	PPVT Stnd	PPVT GSV	CDI Produce	Words PerHour	CTCPerHour	Meaningful	MatEd Ordil
Age	FruitStroop	.196		-.313*	.202	.473**	.457**	.453**	.396**	.410**	.398**	.436**	.012	-.058	-0.094	.288*
	BRIEFGlobal Percentile	.116	-.282*		-.188	-.333**	-.320*	-.312*	-.169	-.167	-.153	-.223	-.148	-.165	-0.179	-.192
	MinPairs	.221	.236	-.156		.259*	.261*	.280*	.213	.221	.216	.216	.156	.227	0.158	.410**
	EVT_Raw	.322*	.503**	-.275*	.310*		.986**	.990**	.764**	.756**	.758**	.587**	.328**	.233	.319*	.282*
	EVT_Stnd	.023	.453**	-.315*	.259*	.941**		.985**	.740**	.743**	.739**	.595**	.319*	.226	.323*	.276*
	EVT_GSV	.319*	.484**	-0.257*	.330**	.991**	.941**		.741**	.737**	.743**	.603**	.323*	.244	.331**	.272*
	PPVT_Raw	.388**	.434**	-.109	.278*	.792**	.690**	.771**		.994**	.995**	.549**	.396**	.291*	.233	.315*
	PPVT_Stnd	.144	.426**	-.147	.246	.755**	.738**	.738**	.962**		.993**	.571**	.387**	.276*	.220	.325*
	PPVT_GSV	.387**	.435**	-.095	.280*	.787**	.690**	.773**	.996**	.962**		.562**	.402**	.313*	.244	.319*
	CDIProduce	.298*	.467**	-.177	.267*	.627**	.575**	.640**	.599**	.583**	.610**		.241	.251*	.229	.089
	WordsPerHour	-.063	.005	-.154	.138	.290*	.317*	.285*	.340**	.373**	.345**	.211		.735**	.742**	.359**
	CTCPerHour	-.069	-.070	-.171	.205	.198	.224	.209	.241	.262*	.262*	.219	.736**		.712**	.215
	Meaningful	.009	-.091	-.177	.156	.305*	.323**	.316*	.218	.219	.228	.222	.740**	.710**		.254*
	MatEdOrdil	-.210	.235	-.211	.345**	.194	.265*	.186	.203	.284*	.207	.020	.364**	.224	.247	

In cases where the child did not attempt the target word, or the response was judged to be unusable (see Stimulus Preparation below), the student testers were instructed to refrain from providing a verbal model or repeating the target prompt. Testers instead used general language such as “what was that?” or “tell me again!” to limit exposure to the target sounds.

2.2 Stimulus preparation

This section describes the process of isolating the target speech productions, annotating the productions, and preparing the productions for use as stimuli in for the perception study.

2.2.1 Recording segmentation

After speech elicitation, a team of trained students isolated and annotated target words in a process referred to as segmentation. Segmentation was performed using Praat software, with scripts written by members of the Learning to Talk team. For each child’s recording, a text grid was created including the target word, boundaries of the production, and production number within a trial. Descriptive notes were included, such as whether the child responded immediately after the target stimulus or if there was intervening speech, and if there were any issues with the recording (e.g. background noise, too quiet or loud). All segmentation text grids were checked by an additional trained student prior to acoustic event tagging.

2.2.2 Acoustic event tagging and stimulus extraction

Acoustic tagging was also performed using Praat software, with custom-written

scripts. The acoustic tagging protocol was based off of that developed by Eunjong Kong and Tim Arbisi-Kelm, and described in Kong and Weismer (2010). A detailed description of the tagging protocol is presented in Appendix A. Two trained graduate students tagged the acoustic events for all recordings. The tagging process consisted of five key elements (Table 8): selecting the production to be tagged, transcribing the initial consonant, noting any atypical characteristics of the production or sound sample, placing a tag for the time of the stop burst ("burst"), and placing a tag for the onset of vocal fold vibration ("VOT").

Table 8: Acoustic event tagging process

Tagging Step	Description	Purpose
1. Select production	When multiple productions of a target were present, the first usable response was selected. Productions were considered unusable if the burst was not audible, the waveform was clipped, or the production was obscured by background noise.	Select the most authentic representation of a talker's target production
2. Transcription	Target sound was assigned a label for manner (stop, affricate, other) and place ([t, t:k, k:t, k], other) based on tagger's perception	Assign a label for production relative to target sound. Used to calculate "accuracy"
3. Add notes	Added notes to textgrid as applicable: background noise, overlapping response, quiet, clipping, deleted, malaprop, short VOT	Document atypical characteristics of a production
4. Tag stop burst	Labeled first clear peak in the waveform, deviating from zero	Identify beginning of consonant production
5. Tag voicing onset	Labeled beginning of quasi-periodic motion in waveform	Identify beginning of vowel production

To begin, the tagger listened to the length of time approximately corresponding to the initial consonant and vowel (as judged by the waveform). The tagger selected this production if a) it was considered to be usable (see below) or b) it was the only

production available, and enough information was present to place tags. If no taggable productions were available for a trial, it would be coded as missing data. Taggers were instructed to always tag the first usable production of a target, meaning that the tagger was able to identify a stop burst and an onset of vocal fold vibration, and there was no noise obscuring the initial consonant and vowel. The first usable production was selected to sample the child's most authentic production ability for a target. Errors in sound production were still considered usable data.

After selecting a production, the tagger transcribed the perceived manner (stop, affricate, other) and place of articulation (alveolar [t], velar [k], intermediate to [t] and [k], Other) of the initial stop (Table 9). Productions transcribed as "affricate" and "Other" for manner, or transcribed as "other" for place of articulation were not presented as perception stimuli.

Table 9: Transcription category descriptions

Transcription	Interpretation
[t]	Perceived as alveolar place of articulation; Counted as "accurate" for target /t/ only
[t:k]	Perceived as intermediate to alveolar and velar place of articulation, but closer to alveolar; Counted as "accurate" for target /t/ only
[k:t]	Perceived as intermediate to velar and alveolar place of articulation, but closer to velar; Counted as "accurate" for target /k/ only
[k]	Perceived as velar place of articulation; Counted as "accurate" for target /k/ only
Other	Perceived as fricative (such as [s]), or place of articulation was outside of the velar to alveolar range (such as [p]); Never counted as "accurate"

Following transcription, the target sequence was annotated to flag any atypical qualities of the production or sound sample. Notes included background noise, overlapping response (child's production overlapped with computer stimulus), quiet (to a degree that a stop burst was not audible), clipping (production was too loud, and peaks of waveform were clipped), deleted (target was attempted but an initial consonant was not produced), malaprop (child produced the target sound, but within a non-elicited word), devoiced vowel, and short VOT (voice onset time of less than 20 msec). No productions were excluded due to use of annotations, but these notes were visible during the process of perception stimulus selection.

Two primary acoustic events were tagged, the stop burst of the initial consonant and the onset of voicing of the following vowel. Stop burst was operationally defined as

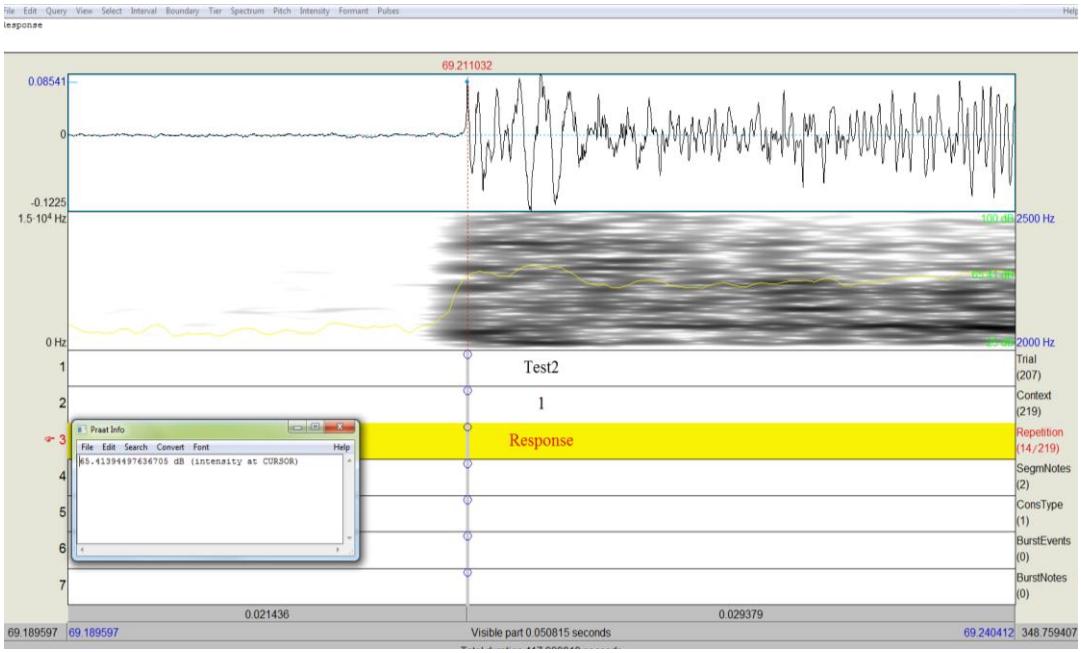

Figure 2: Waveform, spectrogram and textgrid showing cursor location at "burst"

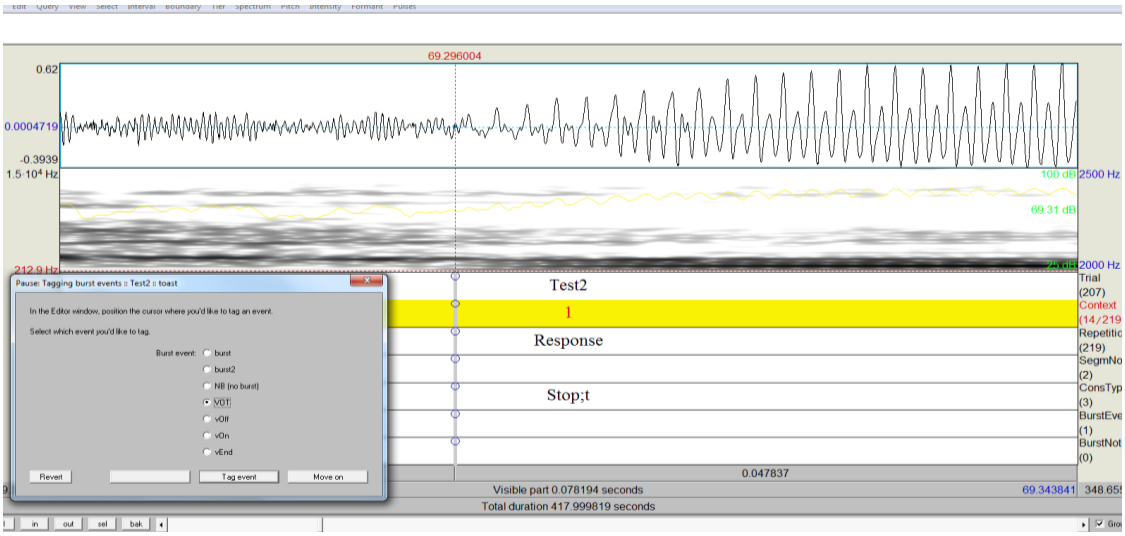

Figure 3: Waveform, spectrogram and textgrid showing cursor location at "VOT"

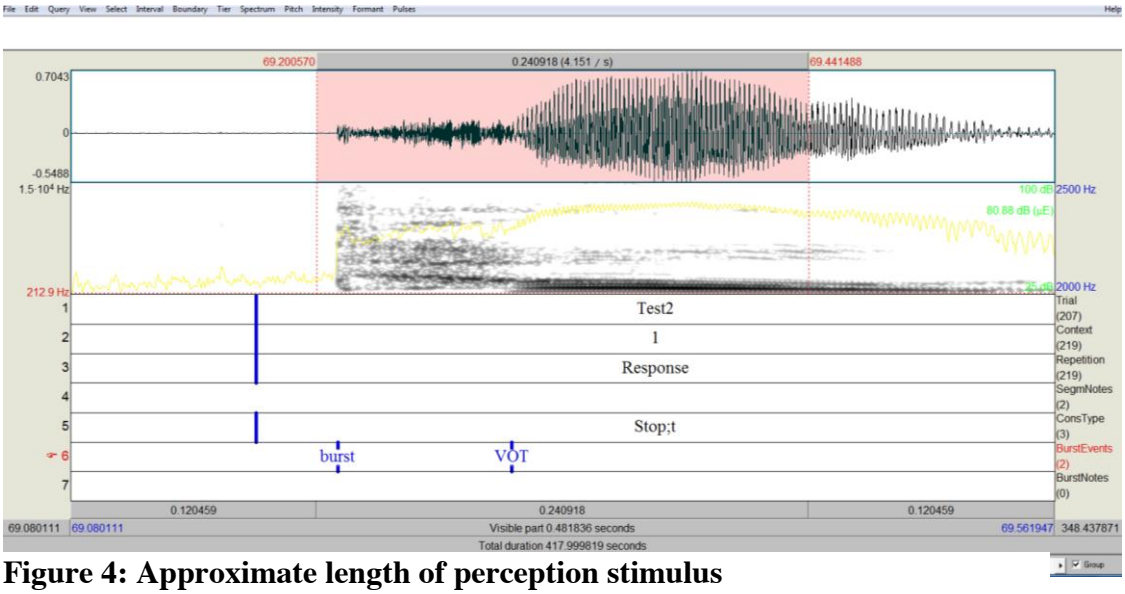

Figure 4: Approximate length of perception stimulus

the first peak of the waveform, as a clear deviation from the baseline waveform of pre-burst lip closure (Figure 2). While tags were primarily placed based on the waveform, presence of energy in the spectrogram was also considered to disambiguate challenging cases. To ensure that noise was not mislabeled as a stop burst, burst peaks were defined as at least 15 decibels (dB) more intense and 20 msec after all other burst candidates. The second tag, voicing onset, was marked at the beginning of the first voicing cycle (Figure

3). This was observed as the first upswing of the cycle prior to first clear downswing below the zero line. Voicing was always tagged at a zero-crossing. Following transcription and event tagging, a compilation of candidate stimuli was created by extracting the audio from 15 msec prior to the burst tag to 150 msec after the VOT tag (Figure 4). Candidate stimuli were those that were transcribed as a stop consonant with both a "burst" and "VOT" tag. Candidate stimuli were excluded from the perception experiment if they were too quiet (burst not audible) or too loud (clipping in waveform), or if background noise occurred during the consonant-vowel sequence, including that from the child's production overlapping the computer prompt. Stimuli were also discarded if the onset of vocal fold vibration fell within 20 msec of the "burst" tag. The stimuli were then amplitude normalized. A total of 1564 productions were prepared for presentation across the three experiment versions (Table 10). All tokens from one talker, 051L, were included in all three experiment versions to be used in intra-rater reliability measures.

Table 10: Number of tokens by transcription categories for unique talkers in versions A, B, C and common talker, 051L

Transcription:	[k] for /k/	[k] for /t/	[k:t]	[t:k]	[t] for /k/	[t] for /t/	total
051L: Common	12 (43%)	1 (4%)	2 (7%)	5 (18%)	0 (0%)	8 (29%)	28
Version A: Unique	200 (38%)	10 (2%)	43 (8%)	34 (7%)	35 (7%)	201 (38%)	523
Version B: Unique	205 (41%)	6 (1%)	35 (7%)	35 (7%)	18 (4%)	201 (40%)	500
Version C: Unique	225 (44%)	9 (2%)	26 (5%)	32 (6%)	17 (3%)	204 (40%)	513

2.3 Adult listeners

This section presents characteristics of the adult listeners and describes the perception study procedure.

2.3.1 Listener participants

The listener participants in this perception study were 47 adults (16 in experiment versions A and B, 15 in version C), tested at the University of Minnesota in Minneapolis (Table 11). A total of 65 listeners was eventually tested for each version; this thesis reports on the listeners whose data had been collected by April 19, 2015. Listeners were recruited via fliers posted around the University campus, in-class announcements to undergraduates in Department of Speech-Language Hearing Sciences lectures, and through word of mouth. All listener participants were self-reported to be native speakers of a North American dialect of English, defined as having acquired a North American dialect of English from birth in North America, from parents who speak North American English. Listeners also had no history of speech, language, or hearing impairments per self-report. Listeners were 18 to 39 years old, and 17 of the 47 listeners were male. A hearing screening was administered at 500, 1000, 2000 and 4000 Hz tones presented at 25 decibels hearing level (dB HL). All but two participants passed the hearing screening (the two listeners did not respond to 4000 Hz at 25 dB HL). Listeners were untrained in rating children's speech.

Table 11: Adult listener information by experiment version

Experiment Version	Number of Listeners	Average Age (years)	Number of females
A	16	22	9
B	16	22	13
C	15	23	8

2.3.2 Perception experiment procedure

This perception experiment was administered on a Dell laptop running E-Prime software. Testing was completed in a quiet room, and stimuli were presented through Sennheiser HD 280 Pro circumaural headphones at a comfortable listening level. A total of 1564 consonant vowel sequences for target /t/ and /k/ were presented to listeners, split into three experiment versions (A, B, C) of roughly equal lengths. The experiment was divided into three versions avoid listener fatigue. Listeners were provided with written and verbal instructions to rate the speech sound along the VAS, with one end labeled 'the "t" sound' and the other end labeled 'the "k" sound' (Figure 5).

Figure 5: VAS presented in perception study

Stimuli were played while the screen showed "Listen". After the stimulus offset, the VAS appeared on the screen. Responses were not timed. Five practice items, representative of both intermediate and non-intermediate productions, from a previously

collected corpus of children's speech samples (paidologos project, Edwards & Beckman, 2008) were presented at the beginning of each experiment version to acclimate listeners to the VAS. Test stimulus presentation order was randomly shuffled for each listener. For the purpose of statistical analysis, click location along the VAS (x-axis click location) was converted so that "the 't' sound" arrowhead = -1.0 and "the 'k' sound" arrowhead = 1.0, with the VAS midpoint = 0.0.

3 Results

The results section is organized as follows. First, listener VAS rating results are presented. This section describes the pattern of results aggregated across listeners, then describes the performance of individual listeners. The next section describes the relationship between the measures of individual differences among the talkers and the listener VAS ratings. That section includes both descriptive correlations and linear regression models.

3.1 Listener results

Listeners were asked to click along the visual analog scale to indicate how /t/- or /k/-like they perceived each token to be. For data analysis, the "t sound" and "k sound" points along the scale were transformed so that "t" = -1.0 and "k"= 1.0, with the line midpoint = 0. The aggregated response clicks were found to be distributed somewhat bimodally along the visual analog scale (Figure 6). This bimodal distribution shows that the listeners had a tendency to perceive sounds as either /t/-like or /k/-like. The fact that the ratings utilized the entire scale indicates that there were numerous sounds perceived

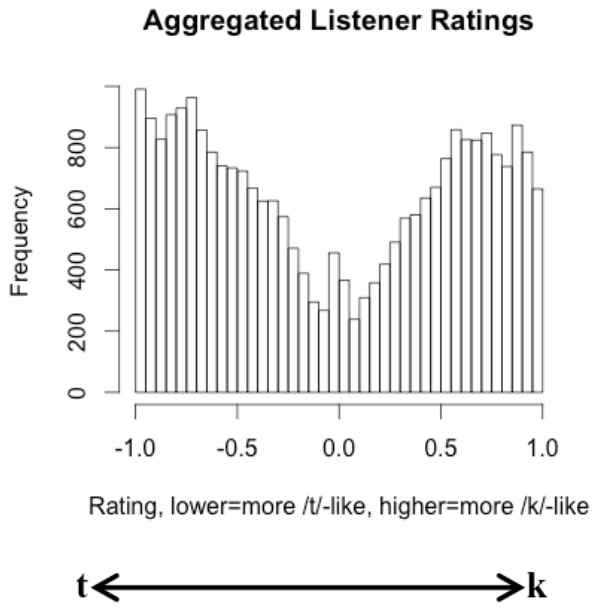

Figure 6: Aggregated ratings along VAS for 47 listeners

as ambiguous or intermediate. This result is what we would predict given that the productions used as stimuli had more instances of sounds transcribed as [t] or [k] than ones transcribed as intermediate.

3.1.1 Aggregated listener differentiation between transcription categories

This section reports on how well listeners, pooled together, were able to differentiate between transcription categories using the VAS. Because the tokens presented in this perception experiment were initially transcribed into four categories ([t, t:k, k:t, k]) listener click location was compared to the trained transcriber's assignment to transcription category. Figure 7 shows boxplots of the click locations, pooled across all listeners, along the visual analog scale for each transcription category. Six categories are shown in this plot because substitutions are different from correct productions.

A linear mixed-effects model was applied to ratings. Normalized click location

(i.e., clicks normalized to the [-1,1] range) was used as the dependent measure. Transcription category was a fixed effect. Terms were added for slopes for individual listeners, and for the effect of transcription category on individual listeners' ratings. A series of models was built, with each of the transcription categories as reference levels. All of these models had a significantly better model fit than a model whose only term was a random intercept for listeners. In all of these models, transcription category had a significant effect on ratings. The only pairwise contrast that was not significant at the a < 0.05 level was that between productions transcribed as [t] for target /t/ and [t] for target /k/. This is somewhat surprising, as this difference has been found previously to differ acoustically. This could be due an oversampling of true [t] for /k/ substitutions rather than covert contrast productions transcribed as [t], as covert contrasts may have been more often transcribed as [t:k]. Comparison of these rating data to acoustic measurements

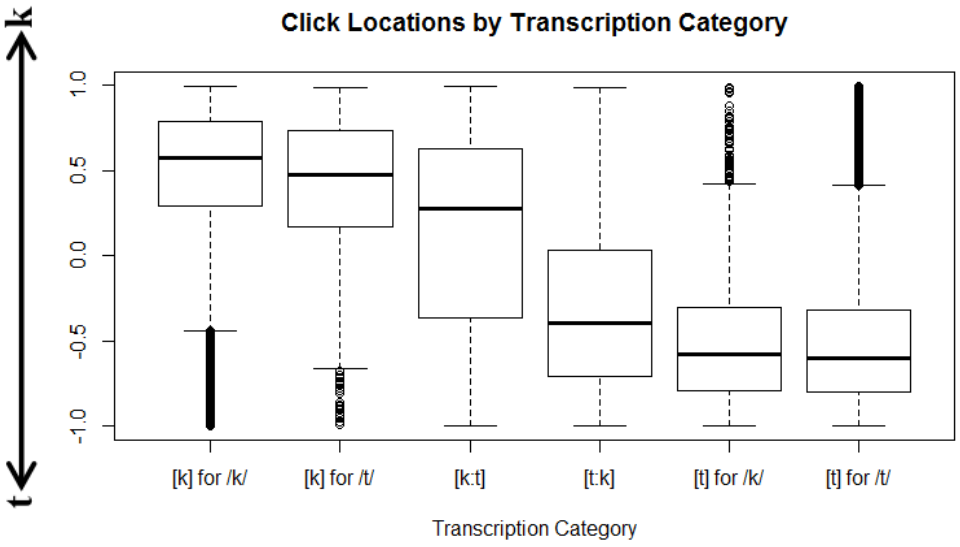

Figure 7: Click location for the six transcription categories

(Johnson, in progress) will provide further illumination regarding the similarities of the [t] for /t/ and [t] for /k/ productions. Another point of interest is the dispersion of rating, as shown by the boxes (i.e., the interquartile range) in Figure 7. The spread in data is significantly greater for the intermediate transcription categories [t:k] and [k:t] than for the [t] and [k] categories. This finding highlights the utility of VAS in more sensitively classifying the differences between intermediate productions than traditional phonetic transcription.

3.1.2 Individual listener differentiation between transcription categories

The next analysis examined the extent to which individual listeners' ratings differentiated among the six types of productions: [t] for /t/, [t] for /k/, [t:k], [k:t], [k] for /t/, and [k] for /k/. To examine this, individual one-way ANOVAs were conducted separately by listeners. Each listener's rating was the dependent measure in his/her own ANOVA, and transcription category was the predictor variable. Post-hoc Scheffe tests were used to derive homogeneous subsets. Homogeneous subsets are subsets of the data which differ from other subsets significantly. They are defined relative to the independent variable. If a listener had perfect differentiation among the six stimulus types, then he/she would have six homogeneous subsets, corresponding to the ratings for the six stimulus types. If a person had only one homogeneous subset, then the person would have no differentiation among the six stimulus types.

All of the individual-subjects ANOVAs were statistically significant, indicating that each listener's ratings discriminated among at least two of the transcription categories. The number of homogeneous subsets varied across was either 2 (14/47

listeners), 3 (26/47 listeners), or 4 (7/47 listeners). The specific categories that were distinguished within the homogeneous subsets varied. In general, most of the two-category listeners' judgments differentiated between sounds transcribed as [t] or [t:k] and those transcribed as [k] or [k:t]. Most of the three-category listeners' judgments differentiated between [t], [k], and intermediate productions. There was no clear pattern for the four-category listeners. This means that for greater than 2/3 of the listeners, the ratings were more informative than a simple two-category transcription system. The distribution of two-, three- and four-category listeners differed significantly across the three versions of the experiment, $\chi^2_{[df=4]}=10.895$, $p=0.028$. There were a greater proportion of two-category listeners in experiment B than in the other two versions. The reason for this is not immediately apparent, and suggests that a more rigorous analysis of the psychometric equivalence of the three versions of the experiment is warranted. It may be that the stimuli in Version B are inherently less ambiguous than those in Versions A and C.

3.1.3 Listener intra-rater reliability

Within each experiment version (A, B, C), 20 productions from different talkers were repeated to gauge intra-rater reliability. For each of the 47 listeners, three measures of intra-rater reliability were obtained: average distance in click location between the two presentations of the same token, correlation between the two click locations, and the percentage of ratings of the 20 repeated tokens that fell within 15% of the entire line distance. For listener reliability measures, the click location along the visual analog scale were transformed to range from 0.0 (the "t" sound) to 1.0 (the "k" sound). Average

distance between click locations along the visual analog scale for sets of repeated tokens ranged from 0.08 to 0.27, mean = 0.14, standard deviation = 0.04. In other words, listeners' two click ratings of the same token were on average within 14% of the total VAS line length apart. Correlation between click ratings of repeated tokens within listeners ranged from 0.47 to 0.97, mean = 0.79, standard deviation = 0.12. A third measure of intra-rater reliability determines what proportion of repeated token pairs was rated within 15% of the VAS line length. This value ranged from 0.45 to 0.90, mean = 0.68, standard deviation = 0.12. On average, only 68% of token pairs were rated within 15% of the VAS line from each other, a somewhat poor level of intra-rater reliability.

It is to be expected that some listeners have higher levels of intra-rater reliability than others. A linear regression model was applied to determine if intra-rater reliability

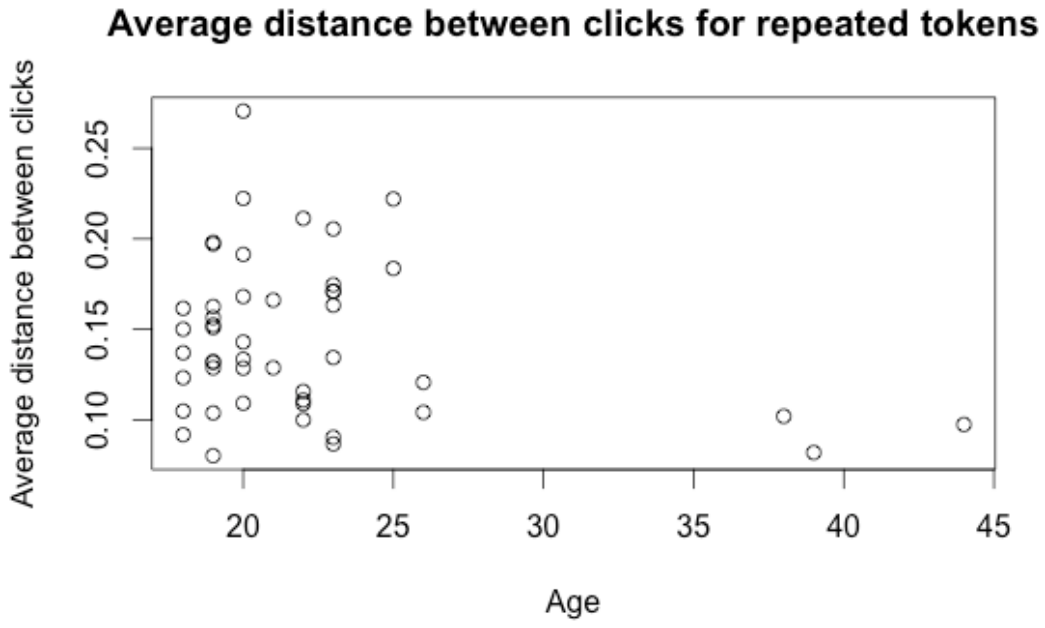

Figure 8: Intra-rater average distance between ratings for repeated tokens

was predicted by experiment version, listener age, or listener sex for each of the three intra-rater reliability measures. Listeners were significantly more reliable on the average distance measure in Version A ($p < 0.01$) than Version B or C. Women were somewhat more reliable than men ($p = 0.055$), and older participants were significantly more reliable than younger participants ($p < 0.05$). Upon inspection of the plots of age against reliability (Figure 8), it was determined that three outliers representing older ages with very low distance between click locations (more reliable) may have been driving this relationship.

To correct for this phenomenon, age was converted to a logarithmic variable (logAge). Using the variable logAge, along with experiment version and sex to predict average difference in click location, logAge was no longer a significant predictor of reliability ($p = 0.068$) while sex approached significance ($p = 0.051$). Correlation between click responses for repeated tokens varied significantly by experiment version. Version A had the highest correlation between clicks, compared to Version B ($p < 0.01$) and Version C ($p < 0.01$). Sex was not a significant predictor of correlation between click locations ($p = 0.24$) while logAge was ($p < 0.05$). The measure of proportion of repeated responses within 15% of the total VAS line length also varied by experiment, although not as strongly as for distance between clicks and correlation between click location. Version A reliability ratings differed from Version C ($p < 0.05$), but not from Version B ($p = 0.13$). Neither sex ($p = 0.57$) nor logAge ($p = 0.09$) predicted this reliability measure.

3.1.4 Set effects

Across the three experiment versions (A, B, C), all productions from one talker,

051L, were included to determine whether set effects were present (Table 10). We were able to determine whether there was a significant difference between the aggregated responses to one talker's speech due to the influence of the other stimuli present in that experiment version. A linear mixed-effects model was applied to see if VAS ratings for 051L varied by experiment version. Overall, there was no main effect for VAS ratings by experiment version. When transcription category was added to the model, an interaction was observed between experiment version and transcription category for VAS ratings. This means that some transcription categories were rated differently depending on the experiment version.

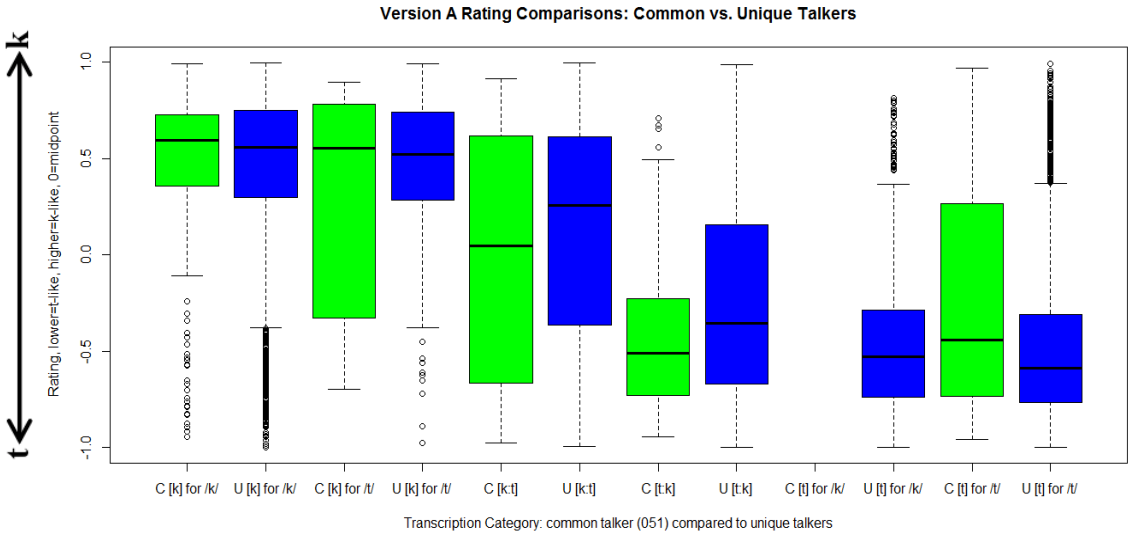

Figure 9a

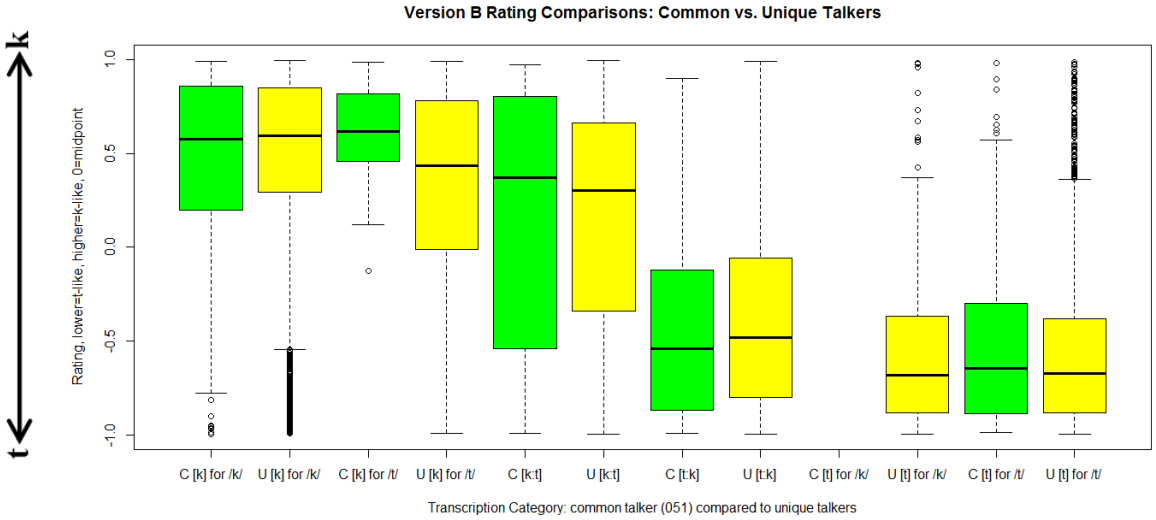

Figure 9b

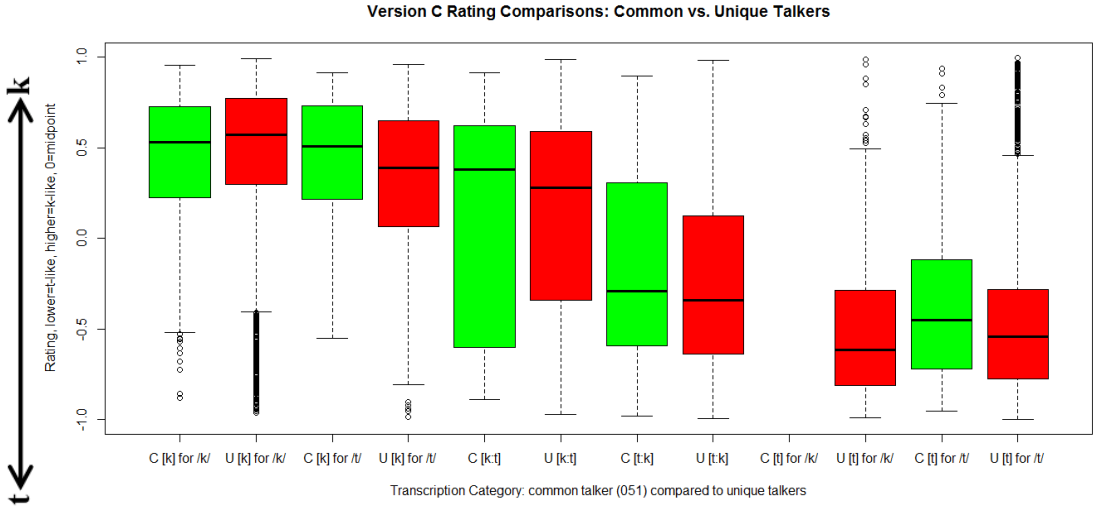

Figure 9c

Figure 9: Ratings of common talker 051L (green, left boxes in pairs) compared to unique talkers (right boxes in pairs) by transcription category

3.2 Talker results

Analysis of the talker-related data, i.e. child-level variables, serves a twofold

purpose. First, child-level factors can determine predicting factors for "accuracy" (table 1) as well as robustness of contrast in /t/-/k/ production to characterize the developmental trajectory of this sound contrast. This comparison is important to illuminate the differences between VAS ratings and transcription accuracy that can validate VAS as meaningful tool in clinical and research purposes. Second, it allows for comparison between predictors of sound accuracy and predictors of robustness of contrast to identify differences in sensitivity between transcription and VAS ratings. In this section, talker-related variables are discussed, the dependent measures of accuracy and slope are explored, and correlations and linear-regression models among child-level variables are presented.

3.2.1 Dependent variables: Slope and accuracy

Two talker-related dependent measures were identified: target accuracy (asinAcc) and slope. The first measure, accuracy was derived from trained transcriber's phonetic transcriptions of the productions. Accuracy was determined by the percentage of a child's productions that were phonetically transcribed within the target stop's category (see table 1). For example, transcriptions of [t] and [t:k] were both counted as accurate for target /t/, but incorrect for target /k/. Likewise transcriptions of [k] and [k:t] were counted as accurate for target /k/, but incorrect for target /t/. Because it has been documented that in proportional scales, variances are correlated with means, and data are not normally distributed around the mean, a rationalized arcsine transform was applied to percent accuracy data (Studebaker, 1985). This transform yielded the dependent measure "asinAcc", which is more suitable for statistical analysis.

$$AU = \sin^{-1}\sqrt{\frac{s}{N+1}} + \sin^{-1}\sqrt{\frac{s+1}{N+1}}$$

then

$$RAU = \left(\frac{146}{\pi}\right) \cdot AU - 23$$

The second dependent measure, slope, was derived from listener VAS ratings to provide a measure of listener-defined robustness of contrast. This measure was calculated for each child by plotting histograms of the listener ratings for attempts at target /t/ and attempts at target /k/ (left panes, Figure 10). The x-axis shows click location along the VAS (-1.0 to 1.0). The bottom histogram in each plot shows where on the VAS the talker's attempts at /k/ were rated, and the top histogram shows where on the VAS the talker's attempts at /t/ were rated. A logistic regression was applied to fit the best curve to these data. The slope of this curve yields the dependent measure "slope". Higher positive slope values represent a more robust contrast between /t/ and /k/ productions, where attempts at /t/ were rated toward the "t" end of the scale and attempts at /k/ were rated toward the "k" end of the scale. Lower values (including negative values) represent a less robust contrast. The y-axis represents the probability that a given click location on the VAS corresponded to an attempt at target /t/. Examples of curves for five children, with a representative range of slope values are presented below (Figure 10). The accuracy measure is not represented graphically in these plots, but included in the plot titles to compare the slope and accuracy measures for each talker. The talkers shown range from least robust contrast (Figure 10a and 10b) to most robust contrast (Figure 10i and 10j) as determined by slope. Robustness of contrast, characterized by the separation between

VAS ratings for productions of target /t/, compared to productions of target /k/, can also be observed within the context of density curves. In these plots, listener VAS ratings are represented along the x-axis while click frequency is shown on the y-axis. Two curves are shown on each density plot, one for productions of target /t/, and one for productions of target /k/. More robust contrasts are represented by curves with less overlapping area between the /t/ and /k/ curves. Side-by-side representations of logistic curves and density curves are presented for each of the five talkers.

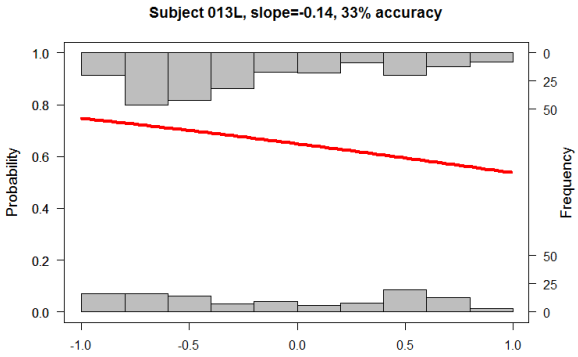

Figure 10a

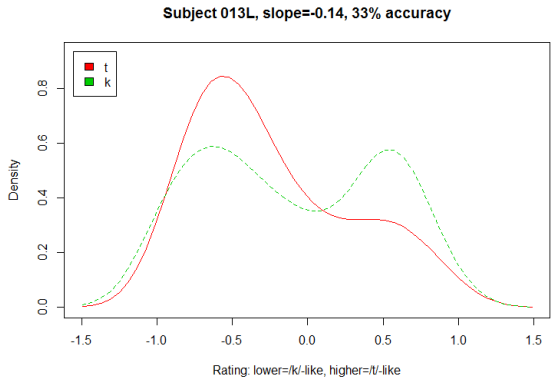

Figure 10b

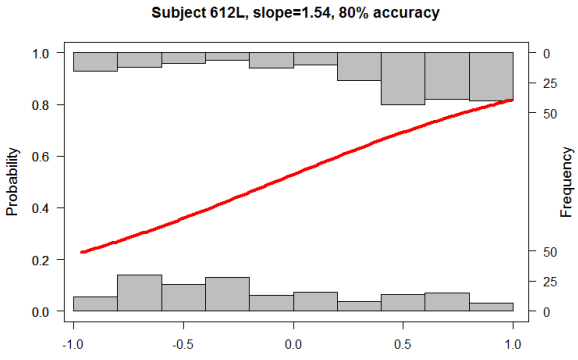

Figure 10c

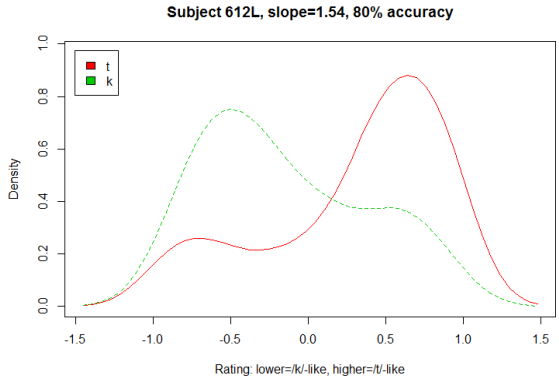

Figure 10d

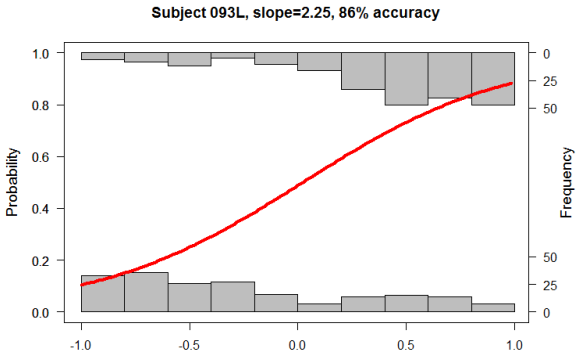

Figure 10e

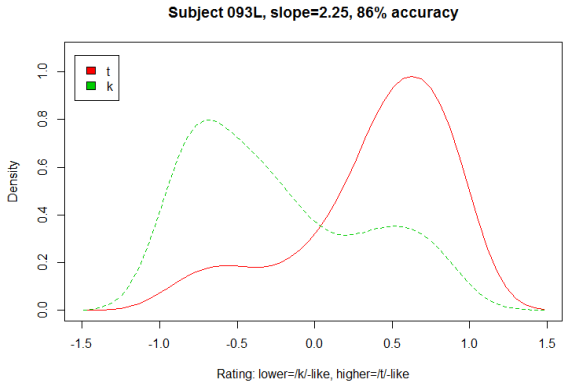

Figure 10f

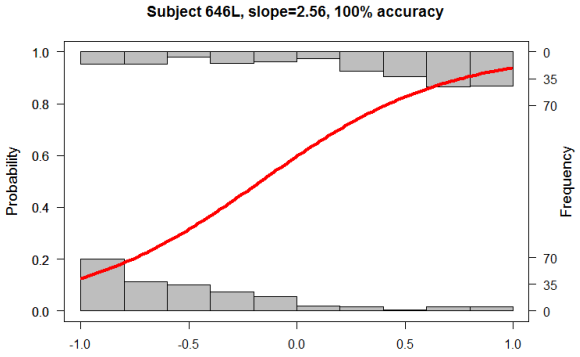

Figure 10g

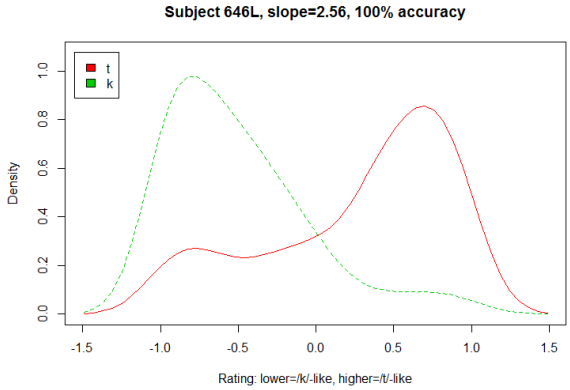

Figure 10h

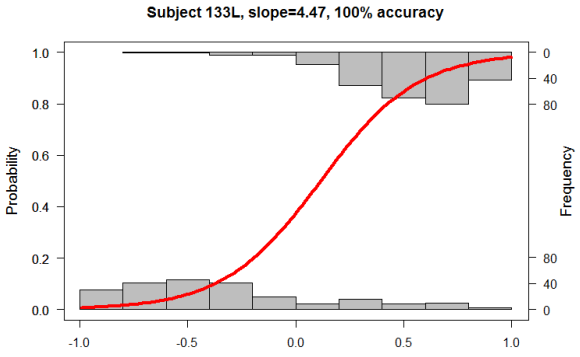

Figure 10i

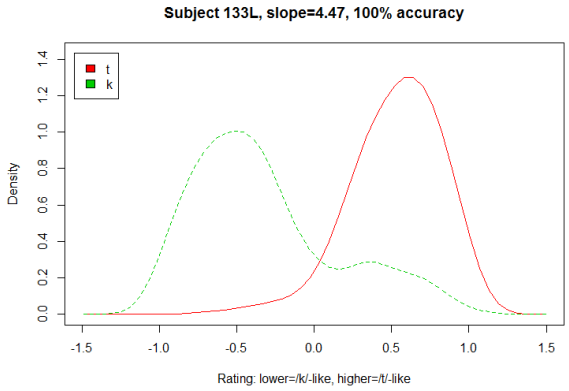

Figure 10j

Figure 10: Side-by-side representations of logistic curves (left) and density plots (right) derived from listener VAS ratings for five children. Talkers shown in top to bottom order of least robust contrast (smallest slope values) to most robust contrast (largest slope values).

Of the two dependent measures, target accuracy as determined by phonetic transcription is the more traditional measure in both research and clinical contexts. However slope, obtained through VAS ratings, is a more granular measure, i.e., one that has more potential values. Slope is thus a potentially more sensitive measure to determine contrast acquisition than accuracy. This can be observed in comparing Figure 10g and 10h (talker 646L) with Figure 10i and 10j (talker 133L). Both talkers were transcribed to 100% accuracy, meaning that transcription category matched the target category for all productions. However talker 133L has a steeper slope (slope = 4.47) than talker 646L (slope = 2.56). Clearly there are differences in listener perception of these two talkers that are overlooked through use of phonetic transcriptions alone.

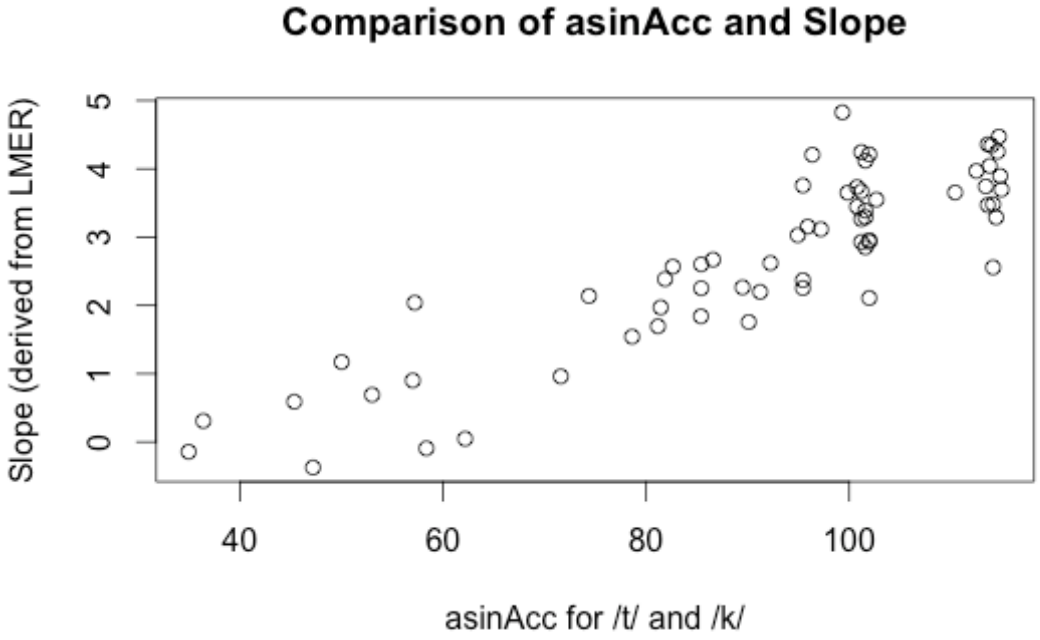

Figure 11: Comparison of asinAcc and listener-derived slope

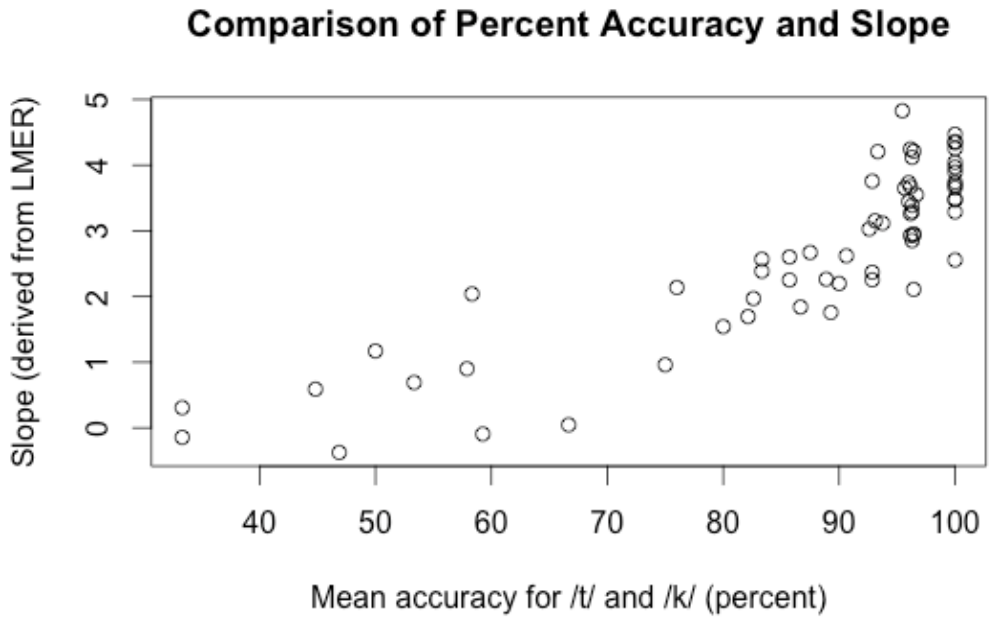

Figure 12: Comparison of transcribed accuracy (percent) and listener-derived slope

Across all talkers, accuracy predicts slope, however these two measures of speech production are nonlinearly related. As shown in Figure 11 and Figure 12, transcribed accuracy is more bounded than slope, condensing the data with higher accuracy. Slope, however, is able to differentiate between the talkers with high transcribed accuracy. These data support the finding that VAS is a tool that is more sensitive to robustness of contrast than traditional phonetic transcription.

3.2.2 Predictors of slope and accuracy

Having identified and established the dependent measures of slope and asinAcc, as well as the input and output independent measures, we can begin to look at what factors predict our two measures of speech production. Descriptive correlations and mixed-effects linear regression models are presented below.

3.2.3 Descriptive correlations

Two tables of descriptive correlations are presented below: Table 12 shows full correlations and Table 13 shows partial correlations, where variables have been residualized for age. In both tables, slope is significantly correlated with age. Additionally, all vocabulary measures (EVT, PPVT, CDI) are significantly correlated with slope and accuracy, however these relationships are stronger in the full correlations than the partial correlations.

Table 12: Full correlations between independent and dependent variables, including the effect of age on each variable. Coefficient estimate is shown, with significance * = p < 0.05, ** = p < 0.01

Full Correlations				
Control Variables		Age	Slope	asin Acc
Age	Slope	-.394**		
	asinAcc	.369**	-.883**	
	FruitStroop	.196	-.206	.172
	BRIEFGlobal Percentile	.116	.056	-.099
	MinPairs	.221	-.065	.040
	EVT_Raw	.322*	-.371**	.378**
	EVT_Stnd	.023	-.277*	.313*
	EVT_GSV	.319*	-.374**	.385**
	PPVT_Raw	.388**	-.405**	.430**
	PPVT_Stnd	.144	-.329**	.368**
	PPVT_GSV	.387**	-.410**	.435**
	CDIProduce	.298*	-.456**	.425**
	WordsPer Hour	-.063	-.006	-.044
	CTCPer Hour	-.069	-.028	-.025
	Meaningful	.009	-.023	-.079
	MatEdOrdil	-.210	-.040	.022

Table 13: Partial correlations between independent and dependent variables, without the effect of age on each variable. Coefficient estimate is shown, with significance * = p < 0.05, ** = p < 0.01

Partial Correlations		
Control Variables	Slope	asinAcc
asinAcc	-0.864**	
Fruit Stroop	-.142	.109
BRIEF Global Percentile	.112	-.153
Min Pairs	.024	-.046
EVT Raw	-.281*	.294*
EVT Stnd	-.292*	.327**
EVT GSV	-.285*	.303*
PPVT Raw	-.298*	.335**
PPVT Stnd	-.300*	.342**
PPVT GSV	-.304*	.341**
CDI Produce	-.386**	.355**
Words PerHour	-.034	-.022
CTCPer Hour	-.060	.001
Meaningful	-0.021	-0.089
MatEdOrdil	-.137	.109

3.2.4 Linear regression models

After inspecting the descriptive correlations, linear regression models (using the lmer package in R software) were analyzed to further determine what independent measures are statistically significant in determining slope and accuracy. Age was kept as a variable in every model, however the independent variables found to be correlated with age were

residualized for the effect of age. The coefficient estimates, t-values, and p-values are provided for the measures found to be significantly correlated with speech production (Table 14).

Table 14: Coefficient estimates and standard error, t-values, and p-values for three linear regression models, using the three predictor variables shown to correlate with speech production, for both slope and asinAcc.

	Slope (derived from LMER)				**Accuracy (asinAcc)**			
	Estimate	Stnd Err	t val	p val	Estimate	Stnd Err	t val	p val
Intercept	2.04	1.38	1.48	0.14	17.80	22.87	0.78	0.44
Age	-0.15	0.04	-3.47	<0.01	2.26	0.67	3.23	<0.01
EVT_GSV	-0.02	0.01	-2.31	0.02	0.43	0.18	2.47	0.02
Intercept	2.04	1.37	1.49	0.14	17.80	22.58	0.79	0.43
Age	-0.15	0.04	-3.29	<0.01	2.26	0.69	3.27	<0.01
PPVT_GSV	-0.02	<0.01	-2.47	0.02	0.42	0.15	2.81	<0.01
Intercept	2.04	1.33	1.54	0.13	17.80	22.44	0.79	0.43
Age	-0.15	0.04	-3.60	<0.01	2.26	0.69	3.29	<0.01
CDI_Produce	<-0.01	<0.01	-3.24	<0.01	0.04	0.02	2.95	<0.01

From the linear regression models, it is clear that all three vocabulary measures (EVT_GSV, PPVT_GSV, and CDI_Produce) are significant in predicting both measures of speech production. The remaining output predictor variables, MinPairs, Fruit Stroop, and BRIEF global percentile, were not significant in predicting speech production. The input related measures (MatEd, WordsPerHour, CTC Per Hour, Meaningful, AWC Percentile) also did not predict speech production measures.

Beyond identifying significant predictor variables for the speech production measures, Table 14 allows for direct comparison between t-values and p-values of the slope measure to those of the asinAcc measure. These values are similar across the models predicting the two different dependent measures, showing slope and asinAcc are

roughly equivalent measures for modeling relationships among components of a child's language and communication system.

4 Discussion

The first aim of this study was to develop and validate a clinical tool for assessing children's /t/-/k/ production that reflects the established gradual nature of contrast acquisition. This was done through collecting VAS ratings from 47 adult naïve listeners, presented with consonant-vowel sequences produced by 63 children representing a range in language-related skills and measures. A measure of listener-defined robustness of contrast, slope, was compared to the phonetic transcriptions assigned by trained transcribers. Aggregated VAS ratings demonstrated that listeners were able to differentiate the following transcription categories using a VAS: [k] for /k/, [k] for /t/, [k:t], and [t:k]. Listeners differentiated [t:k] from [t], however rated the [t] for /t/ productions similarly to the [t] for /k/ productions, perhaps due to oversampling of true substitutions. Ratings for the intermediate categories, [k:t] and [t:k], were found to be more distributed along the VAS than non-intermediate categories. This finding supports the claim that VAS lends a specificity in rating sounds which phonetic transcription is not able to provide. Individual listener ratings were analyzed, and greater than 2/3 of listeners were able to differentiate between at least three transcription categories. For these listeners, VAS ratings were more informative than a two-alternative forced choice rating system.

The methodological question of intra-rater reliability was also addressed. Three measures of intra-rater reliability were described: average distance in click location

between repeated tokens, correlation between click locations of repeated tokens, and proportion of repeated tokens that were rated within 15% of the total VAS length from each other. Listener age and sex contributed somewhat to predicting the different intra-rater reliability measures, although not in consistent ways across the three measures. Overall, listeners had poor to fair intra-rater reliability. Currently, no standard for determining a listener to be "reliable" exists. These data will help us develop a better idea of what counts as a reliable listener and what characteristics contribute to the likelihood of a listener counting as reliable. Set effects were pervasive in analysis of listener data. The questions of listener reliability and influence of surrounding stimuli will continue to hold great importance as VAS rating becomes more widespread in clinical and research environments.

The second aim of this study was to identify child-level predicting factors for differentiation of similar articulatory gestures for /t/ and /k/. This was done through collection of a host of output (vocabulary, executive function, speech perception) and input (home language environment, maternal education, late talker status, dialect) measures. Through descriptive correlations and linear regression models, vocabulary size (measured via EVT, PPVT, CDI) was determined to be the only significant predictor of the speech production measures. This finding supports other reports of the relationship between vocabulary size and phonological knowledge in the literature (Edwards, Beckman & Munson, 2004). However, this study and work by Nicholson (2014) are the first to demonstrate the effect of vocabulary size on speech production rather than higher-level phonological knowledge.

Additionally, this thesis described the listener-defined robustness of contrast measure, slope, and compared it to the more traditional measure of transcribed accuracy. As dependent variables, slope and asinAcc are predicted similarly in linear regression models with both input and output independent variables. This occurrence raises the questions of whether slope is just a more time consuming and difficult to compute measure of accuracy. However, slope is clearly informative to differentiate among speech production abilities of talkers with high transcription accuracy. Slope describes the degree in overlap of listener perception between productions for contrasting targets with much finer granularity than accuracy. This is especially true for talkers with high (approximately greater than 85%) transcribed accuracy.

4.1 Contributions to the literature

This study provides support to the growing body of evidence that speech sound contrasts are acquired gradually. Further, sound productions contain more information about a child's phonological output knowledge than phonetic transcription can encode. Therefore, rating speech sounds along a VAS is a more appropriate and informative measure of speech production than phonetic transcription in both clinical and research environments. This study provided an introductory perspective into the methodological factors in utilizing VAS, including intra-rater reliability and the influence of surrounding stimuli on perception. Finally, this study built upon evidence to establish vocabulary size as a key predictor in speech sound differentiation for the /t/-/k/ contrast.

4.2 Limitations

This study contained several limitations. First, the number of listeners in each experiment version, off of which all perception ratings were analyzed, was limited (15 to 16 listeners per experiment version). Additionally, some listeners had exposure to languages other than English, or did not pass a hearing screening at all frequencies presented. At the writing of this thesis, perception testing is ongoing to collect the ratings of at least 20 listeners in each experiment version, who more fully meet the inclusion criteria of this study. Additionally, strong set effects were observed on VAS ratings and intra-rater reliability. The set effects indicate that tokens included in the three experiment versions may not have been distributed in a balanced manner.

4.3 Future directions

This introduces many avenues of further exploration. Future directions are suggested for both listener- and talker-related questions.

4.3.1 Listeners

Future directions to this research should investigate the clinical and research significance of the different intra-rater reliability measures. Studies should examine further what characteristics predict listener reliability, and whether reliability can be trained. These studies could identify the smallest set of listeners needed to get reliable VAS ratings. Another route for future perception studies would be to examine whether VAS ratings of trained listeners (such as speech-language pathologists experienced in judging children's speech accuracy) would prove to be more informative in models

predicting speech production than those of naïve listeners.

4.3.2 Talkers

Relationships among the input and output predictor variables merit further investigation to better characterize children's language development. Future studies should continue to model a variety of speech and language skills, including measures of vocabulary size. These studies can begin to answer questions regarding the effect of vocabulary interventions for children with speech and language delays and disorders. Additionally, the specific clinical implications of the slope value warrant further investigation.

The input predictor variables of status as a late talker and dialect were not fully explored in this thesis. Further work should explore whether these variables predict measures of speech prediction beyond the contributions of age and vocabulary size.

A large number of additional talkers have been recorded as part of the Learning to Talk project. Future perception studies should include speech productions from this larger set of talkers. A number of parallel perception studies involving other sound contrasts (/s/-/ʃ/, /d/-/g/) are in progress by this group of researchers. Comparing results of all perception studies using the same set of talkers and predictor measures will allow us to determine if certain child-level factors are more predictive of one sound contrast, more than the other contrasts. Finally, acoustic analyses on all productions included in this study are currently being performed as the focus of Johnson (in progress). Identifying specific acoustic markers corresponding to the different transcription categories and VAS ratings will contribute to the body of knowledge on children's speech development.

5 Bibliography

Archibald, S. J., & Kerns, K. A. (1999). Identification and description of new tests of executive functioning in children. Child Neuropsychology, 5(2), 115-129.

Babel, M. E., & Munson B. (2013). Producing socially meaningful linguistic variation. (M. Goldrick, V. Ferreira, M. Miozzo, Ed.).Oxford Handbook of Language Production.

Beckman, M. E., Munson, B., & Edwards, J. (2014). Effects of speaker language and listener language on children's stop place. Presentation at The 14th Conference on Laboratory Phonology, Tachikawa, Japan.

Bleile, K. (2002). Evaluating articulation and phonological disorders when the clock is running. *American Journal of Speech-Language Pathology*, 11(3), 243-249.

DeLoach, L. J., Higgins, M. S., Caplan, A. B., & Stiff, J. L. (1998). The visual analog scale in the immediate postoperative period: intrasubject variability and correlation with a numeric scale. Anesthesia & Analgesia, 86(1), 102-106.

Dunn, L. M., & Dunn, D. M. (2007). Peabody Picture Vocabulary Test (4th ed.). Minnesota: Pearson.

Edwards, J. R., & Beckman M. E. (2008). Methodological questions in studying consonant acquisition. Clinical Linguistics & Phonetics. 22(12), 937-56.

Edwards, J., Beckman, M. E., & Munson, B. (2004). The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *Journal of Speech, Language, and Hearing Research*, *47*(2), 421-436.

Edwards, J., Gibbon, F., & Fourakis, M. (1997). On discrete changes in the acquisition of the alveolar/velar stop consonant contrast. Language and Speech, 40(2), 203-210.

Edwards, J., Munson, B., & Beckman, M. E. (2011). Lexicon–phonology relationships and dynamics of early language development–a commentary on Stoel-Gammon's 'Relationships between lexical and phonological development in young children'. Journal of child language, 38(01), 35-40.

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. H., Reznick, J. S., & Bates, E. (2007). MacArthur-Bates Communicative Development Inventories: User's guide and technical manual (2nd ed.). Maryland: Paul H. Brookes Publishing Co.

Forrest, K., Weismer, G., Elbert, M., & Dinnsen, D. A. (1994). Spectral analysis of

target-appropriate/t/and/k/produced by phonologically disordered and normally articulating children. *Clinical linguistics & phonetics*, *8*(4), 267-281.

Forrest, K., Weismer, G., Hodge, M., Dinnsen, D. A., & Elbert, M. (1990). Statistical analysis of word-initial/k/and/t/produced by normal and phonologically disordered children. *Clinical Linguistics & Phonetics*, *4*(4), 327-340.

Gardner, H. (1997). Are your minimal pairs too neat? The dangers of phonemicisation in phonology therapy. International Journal of Language & Communication Disorders, 32(2s), 167-175.

Gibbon, F. (1990). Lingual activity in two speech-disordered children's attempts to produce velar and alveolar stop consonants: evidence from electropalatographic (EPG) data. International Journal of Language & Communication Disorders, 25(3), 329-340.

Gierut, J. A., & Dinnsen, D. A. (1986). On word-initial voicing: Converging sources of evidence in phonologically disordered speech. Language and Speech, 29(2), 97-114.

Gioia, G.A., Espy, K.A., & Isquith, P. K. (2003). Behavior Rating Inventory of Executive Function–Preschool Version: Professional manual. Florida: PAR.

Gilkerson, J., & Richards, J. A. (2009). The power of talk: Impact of adult talk, conversational turns, and TV during the critical 0-4 years of child development (2nd ed.). Retrieved from http://www.lenafoundation.org/wp-content/uploads/2014/10/LTR-01-2_PowerOfTalk.pdf

Holliday, R., Reidy, P., Beckman, M., Edwards, J. (2014). Quantifying the robustness of English sibilant contrast in children. *Journal of Speech, Language, and Hearing Research* (Submitted).

Julien, H. M., & Munson, B. (2012). Modifying speech to children based on their perceived phonetic accuracy. Journal of Speech, Language, and Hearing Research, 55(6), 1836-1849.

Khan, L. M. (2002). The Sixth ViewAssessing Preschoolers' Articulation and Phonology From the Trenches. American Journal of Speech-Language Pathology, 11(3), 250-254.

Kong, E. J., & Weismer G. (2010). Correlation of acoustic cues in stop productions of Korean and English adults and children. Journal of the Korean Society of Speech Sciences. 2(4), 29-37.

Ladd, D. R. (2011). Phonetics in phonology. The Handbook of Phonological Theory, Second Edition, 348-373.

Li, F. (2012). Language-specific developmental differences in speech production: A cross-language acoustic study. Child Development. 83(4), 1303-1315.

Locke, J. L. (1980). The Inference of Speech Perception in the Phonologically Disordered Child. Part IISome Clinically Novel Procedures, Their Use, Some Findings. *Journal of Speech and Hearing Disorders*, *45*(4), 445-468.

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(01), 41-74.

MacLeod, A. A., & Glaspey, A. M. (2014). A multidimensional view of gradient change in velar acquisition in three-year-olds receiving phonological treatment. *Clinical linguistics & phonetics*, *28*(9), 664-681.

Marin, S., Pouplier, M., & Harrington, J. (2010). Acoustic consequences of articulatory variability during productions of /t/ and /k/ and its implications for speech error research. The Journal of the Acoustical Society of America, (1), 445.

Mayo, C., Gibbon, F., & Clark, R. A. (2013). Phonetically Trained and Untrained Adults' Transcription of Place of Articulation for Intervocalic Lingual Stops With Intermediate Acoustic Cues. *Journal of Speech, Language, and Hearing Research*, *56*(3), 779-791.

McCormack, J., McLeod, S., McAllister, L., & Harrison, L. J. (2009). A systematic review of the association between childhood speech impairment and participation across the lifespan. International Journal of Speech-Language Pathology, 11(2), 155-170.

Munson, B., Edwards, J., & Beckman, M. E. (2005). Relationships between nonword repetition accuracy and other measures of linguistic development in children with phonological disorders. *Journal of Speech, Language, and Hearing Research*, *48*(1), 61-78.

Munson, B., Edwards, J., Schellinger, S. K., Beckman, M. E., & Meyer, M. K. (2010). Deconstructing phonetic transcription: Covert contrast, perceptual bias, and an extraterrestrial view of Vox Humana. Clinical linguistics & phonetics, 24(4-5), 245-260.

Munson, B., Johnson, J. M., & Edwards, J. (2012). The Role of Experience in the Perception of Phonetic Detail in Children's Speech: A Comparison Between

Speech-Language Pathologists and Clinically Untrained Listeners. *American Journal of Speech-Language Pathology*, *21*(2), 124-139.

Nicholson, H. B. M. (2014). Exploring variation in accuracy and contrast for sibilant fricatives at the onset of fricative acquisition. Unpublished MA Thesis, University of Minnesota

Nittrouer, S. (1996). The Relation Between Speech Perception and Phonemic Awareness Evidence From Low-SES Children and Children With Chronic OM. *Journal of Speech, Language, and Hearing Research, 39*, 1059-1070.

Perkell, J. S., Matthies, M. L., Tiede, M., Lane, H., Zandipour, M., Marrone, N., Stockman, 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.

Plummer, A. R., Munson B., Ménard L., & Beckman M. E. (2013). Examining the relationship between the interpretation of age and gender across languages. 21st International Congress on Acoustics, 165th Meeting of the Acoustical Society of America, 52nd Meeting of the Canadian Acoustical Society. 19, 060080.

Romeo, R., Hazan, V., & Pettinato, M. (2013). Developmental and gender-related trends of intra-talker variability in consonant production. *The Journal of the Acoustical Society of America*, *134*(5), 3781-3792.

Rvachew, S., & Brosseau-Lapré, F. (2012). Developmental phonological disorders: Foundations of clinical practice. Plural Pub.

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.

Stoel-Gammon, C. (1991). Normal and disordered phonology in two-year-olds. *Topics in language disorders*, *11*(4), 21-32.

Strömbergsson, S. (2014). The/k/s, the/t/s, and the inbetweens: Novel approaches to examining the perceptual consequences of misarticulated speech. Unpublished PhD Thesis, KTH, Stockholm.

Studebaker, G. A. (1985). A "rationalized" arcsine transform. *Journal of Speech and Hearing Research, 28*, 455-462.

Tyler, A. A., & Tolbert, L. C. (2002). Speech-language assessment in the clinical setting.

American Journal of Speech-Language Pathology, 11(3), 215-220.

Tyler, A. A., Figurski, G. R., & Langsdale, T. (1993). Relationships between acoustically determined knowledge of stop place and voicing contrasts and phonological treatment progress. *Journal of Speech, Language, and Hearing Research, 36*(4), 746-759.

Urberg-Carlson, K., Munson, B., & Kaiser, E. (2009). Gradient measures of children's speech production: Visual analog scale and equal appearing interval scale measures of fricative goodness. *The Journal of the Acoustical Society of America*, *125*(4), 2529-2529.

Walker, D., Greenwood, C., Hart, B., & Carta, J. (1994). Prediction of school outcomes based on early language production and socioeconomic factors. Child development, 65(2), 606-621.

Williams, K.T. (2007). Expressive Vocabulary Test (2nd ed.). Minnesota: Pearson.

6 Appendix A: Burst tagging manual

Purpose

The purpose of burst tagging is to identify and label the exact point of the stop burst release in word-initial velar (/k,g/) and alveolar (/t,d/) stops. Before tagging burst events, the following steps have taken place: child was recorded during real word repetition task, segmentation of word repetition recording has been performed, segmentation has been checked. After burst events have been tagged, the following activities may proceed: acoustic analysis of the burst window (5ms before the burst tag to 20ms after the burst tag), extraction for use as stimuli in a perception experiment (15ms before the burst tag to 150ms after the VOT tag).

Manual edited by Sara Bernstein and Allie Johnson in Spring 2015, adapted from original manual by Eunjong Kong and Tim Arbisi-Kelm.

Burst tagging is performed in Praat software using custom-written scripts developed by the Learning to Talk team.

Praat Settings

a. Set the dynamic range in Praat to 40 dB from the menu "Spectrum->Spectrogram settings..."
b. Set the pitch range in Praat to 2000-2500 Hz from the menu "Pitch->Pitch settings..."
c. From the Intensity drop-down menu, make sure the setting "Show intensity" is engaged.
d. From the Intensity drop-down menu, click “Intensity settings” and make sure the view range is 25 to 100 dB.

Components of the tagging script

Select a response to tag: Always select the first usable response (not overlapping with computer prompt, no clipping, no background noise, audible burst)
Select the manner: Options are Stop, Affricate, Other (See section "Perceptual Judgment" below)
Select the place: Will be somewhere along the /t/-/k/ or /d/-/g/ continuum or "other" (See section "Perceptual Judgment" below)
Add notes to the BurstNotes tier (See sections "Notes Tier" below)
Mark release of stop burst (See section "Tagging burst" below)
Mark onset of vocal fold vibration, referred to in this context as "VOT" (See section "Tagging VOT" below)
Perceptual Judgment

a) Decide on the "Consonant type", choosing from among:
optionMenu("Consonant type", 1)
option("Stop")
option("Affricate")
option("Other")
option("NoResponseisTaggable")

b) If it's tagged as consonant_type=="Stop" and the target is /t/, then choose from among:
optionMenu("Stop place", 1)
option("t")
option("t:$k")
option("$k:t")
option("$k")
option("other")
or if it's tagged as consonant_type=="Stop" and the target is /k/, then choose from among:
optionMenu("Stop place", 4)
option("$t")
option("$t:k")
option("k:$t")
option("k")
option("other")

c) If consonant_type=="Stop" or consonant_type=="Affricate" then tag the following events:
burst
VOT
Tagging "burst"

Tagging the Burst

Locate the burst onset in the waveform, mark the burst onset by clicking in the corresponding location in the 'event' point tier, and then click 'Continue'
a. Definitions:
- burst = the "clump" or "clumps" of spikes that make up the transient of constriction release.
- burst onset = the peak of the individual spike that is selected and marked to denote the beginning of the burst. Criteria for choosing this are presented below.
-*peak*= a single spike or peak within the burst, which may or may not also represent the burst onset
b. When one burst is present:
- Find and mark the first peak, which is represented by the first clear deviation from the baseline waveform of the pre-burst closure (this peak can be either positive or negative amplitude).

- When the first peak is of questionable size (e.g., is followed by a much larger peak), find the intensity level of both peaks by placing the cursor at each peak and pressing <F8>. If the first peak is within 15 dB of the larger peak, then select the first spike as the burst; otherwise, mark the second peak as the burst.
c. When two bursts are present: add protocol for double burst
- First measure and compare the intensity of one of the highest-amplitude peaks within each of the two bursts. Are the intensity levels of these peaks within 15 dB of each other?
- No= select the burst containing the peak with greater intensity, and follow instructions in part a. above.
- Yes= are the peaks within 20 ms of each other?
- Yes= select the first burst, and follow instructions in part b.
- No= select the second burst, and follow instructions in part b.
- NOTE: When bursts are more than 20ms apart, this first burst is often either the result of lip opening or a background noise. Further evidence indicating this will be an absence of frication between the two bursts: multiple bursts resulting from the constriction release will almost always be separated by slight frication, which continues after the final burst until it is replaced by aspiration of the wider aperture directly preceding voicing.
d. When three or more bursts are present:
- Again, first measure and compare the intensity of one of the highest amplitude peaks within each of the three bursts. After identifying the burst with the absolute highest amplitude peak of the three, select the earliest burst whose highest-amplitude peak does not measure below 15 dB of the absolute highest amplitude peak. If two adjacent burst candidates (i.e., with no intervening burst candidates) are more than 25 ms apart, follow the instructions above in part c.
- NOTE: Recordings with a moderate level of background noise will sometimes render the intensity comparison uninformative. When this is the case, consult the spectrogram for burst evidence within a darker energy band (i.e., higher amplitude) spread across a relatively wide frequency range, and rely more heavily on the distance criterion (i.e., within 25 ms).
e. When no burst is present:
- Mark the burst onset at the point where the energy begins (e.g., frication resulting from incomplete closure) in the 'event' tier.
- Label the event 'NB' (i.e., "no burst") in the 'eventNote' tier.

Tagging "VOT"

Locate the voicing onset in the waveform, mark it by clicking in the corresponding location in the 'event' point tier, and then click 'Continue'.
a. Looking at the waveform and scanning rightward, locate the beginning of the first voicing cycle, indicated by an upward swing rising above the zero point. It also may help starting from the vowel and scanning leftward, until the point where the waveform becomes periodic (and more sinusoidal-looking).
b. Often this upward deviation from the zero point is very subtle, and followed by a steep fall below the zero point.

c. Place the cursor as closely as possible to this point. Although the script will automatically move the cursor to the zero-crossing after you click 'Continue', you can do this manually by pressing <Ctrl+0> if you wish to test how close your marking is to the zero-crossing.
d. Look at the spectrogram to see if this point aligns adequately with the voicing bar.
- The 'voicing bar' is a row of striated energy in the very low frequencies, corresponding to the energy in the first and second harmonics (typically the strongest harmonics in speech). For men, this is about 100-150 Hz, while for women it can be anywhere between 150-250 Hz, and of course there is lots of variation both within and between individuals.
- If the upswing zero-point occurs much earlier than the voicing bar evidence, mark instead the next zero-point upswing to the right, even if it occurs after an initial downswing.
e. prevoicing: when voicing begins before the burst
- Mark the VOT at the beginning of the first voicing cycle, which now occurs before the first burst.
- Directly beneath the point where you marked vot1, manually type in "pre-voicing" in the eventNote tier.
- If the pre-voicing is not sustained--i.e., stops and then starts again--then in the event tier manually add the label 'vOff' at the point where voicing stops, and then 'vOn' at the point where voicing begins again.
f. devoicing:
- when there is partial devoicing, the waveform becomes slightly aperiodic, making it more difficult to isolate the voicing onset. In these cases there is no absolute "upswing" in the waveform to indicate the initiation of the voicing cycle; however, this complex waveform will still maintain an overall sinusoidal shape, and therefore the VOT should be marked at the first upswing of this "global rise". The voicing bar in the spectrogram should also be relied upon more heavily to locate VOT in these devoiced cases.
- when there is complete devoicing, with no evidence of any periodicity in the waveform (or voicing striations in the spectrogram), then do not label VOT.

Notes Tier

Notes write to the BurstNotes tier. They are visible to stimulus selectors for perception experiments. Notes help other people looking at the textgrid, or yourself at a later time, understand why you selected a certain consonant type or place, why you labeled something as missing data, or any reservations you have about the usefulness of a certain production.
Quiet: Burst is not audible or extremely soft, or signal to noise ratio between background noise and burst is very poor
Clipping: Peaks of the waveform are clipped, may also appear as striation in spectrogram
BackgroundNoise: Could be a transient, talking, rattle, microphone noise, etc. It is especially important to note BackgroundNoise if the noise occurs within the perception stimulus window (15ms before burst to 150ms after VOT)
Short VOT: refers to cases where the onset of vocal fold vibration falls within the burst

analysis window, i.e. the "VOT" tag is LESS THAN 20ms after the "burst" tag. Technically, the stimulus preparation scripts will automatically pass over these stimuli with or without the "Short VOT" tag, but it is useful to have it in the textgrid.
OverlappingResponse: This refers to cases when the child begins
Devoiced vowel: This tag should be used when the vowel is completely devoiced/whispered. When you use the "Devoiced vowel" note, you should not mark a VOT because there is no onset of vocal fold vibration. Sometimes the vowel is partially devoiced, in which case you do NOT use this note, and just place the "VOT" tag where the vocal folds begin vibration. The "Devoiced vowel" note is only for fully devoiced vowels.
Deleted: Use the "Deleted" tag when the child omits the initial stop consonant, for example "at" for "cat". In these cases, tag manner as "Other" (which means you do not select a place of articulation) and select the "Deleted" note.
Additional notes: You may also add in any text you would like in the field "AdditionalNotes". These could include marking when you are unsure how to tag a production