HHS Public Access

Author manuscript

Dev Psychol. Author manuscript; available in PMC 2019 August 01.

Published in final edited form as:

Dev Psychol. 2018 August ; 54(8): 1472–1491. doi:10.1037/dev0000542.

Speech categorization develops slowly through adolescence

Bob McMurray,

Dept. of Psychological and Brain Sciences, Dept. of Communication Sciences and Disorders, Dept. of Linguistics, Dept. of Otolaryngology, University of Iowa

Ani Danelz,

Dept. of Communication Sciences and Disorders, University of Iowa

Hannah Rigler, and

Dept. of Psychology, University of Iowa

Michael Seedorff

Dept. of Biostatistics, University of Iowa

Abstract

The development of the ability to categorize speech sounds is often viewed as occurring primarily during infancy via perceptual learning mechanisms. However, a number of studies suggest that even after infancy, children's categories become more categorical and well-defined through about age 12. We investigated the cognitive changes that may be responsible for such development using a visual world paradigm experiment based on (McMurray, Tanenhaus, & Aslin, 2002). Children from three age groups (7–8, 12–13, and 17–18 years) heard a token from either a b/p or s/ʃ continua spanning two words (*beach/peach, ship/sip*), and selected its referent from a screen containing four pictures of potential lexical candidates. Eye-movements to each object were monitored as a measure of how strongly children were committing to each candidate as perception unfolds in real-time. Results showed an ongoing sharpening of speech categories through 18, which was particularly apparent during the early stages of real-time perception. When analysis targeted to specifically within-category sensitivity to continuous detail, children exhibited increasingly gradient categories over development, suggesting that increasing sensitivity to fine-grained detail in the signal enables these more discrete categorization. Together these suggest that speech development is a protracted process in which children's increasing sensitivity to within-category detail in the signal enables increasingly sharp phonetic categories.

Keywords

Speech Perception; Real-time Processing; Eye-Tracking; Development; Adolescence

Corresponding Author: Bob McMurray, Dept. of Psychological and Brain Sciences, E11 SSH, University of Iowa, Iowa City, IA 52245, 319-335-2408 (voice), 319-335-0191 (fax), bob-mcmurray@uiowa.edu.

Introduction

Speech perception is a complex problem that skilled listeners overcome effortlessly. To identify phoneme categories, listeners must integrate dozens of transient acoustic cues and cope with variability due to talker, coarticulation and speaking rate. Infants and children face an additional and more difficult problem: they must learn the sound categories and words of their language while simultaneous sly confronting the difficult perceptual problem faced by adults.

The canonical view is that children acquire the speech categories of their language early, by around 24 months (Kuhl, Conboy, Padden, Nelson, & Pruitt, 2005; Werker & Curtin, 2005). However, more recent studies suggest development throughout childhood (Hazan & Barrett, 2000; Nittrouer, 2002; Nittrouer & Miller, 1997; Slawinski & Fitzgerald, 1998). These studies leave open the question of how late this development continues, and what aspects of speech perception are still developing. These studies appear to show children forming more discrete categories over development. Yet, this conflicts with the consensus that adult speech categories are gradient, and this gradiency is helpful for perception (Andruski, Blumstein, & Burton, 1994; McMurray et al., 2002; Miller, 1997). The present study extends this investigation of older children's speech perception to understand precisely which aspects of speech perception develop. Given the age at which these developments are occurring, this has important implications for the kinds of developmental processes involved in acquiring the phonology of language.

The Development of Speech Categorization

An early developmental problem in language acquisition is determining the number and structure of the phonological categories in the language. English learning babies must learn to distinguish /r/ and /l/, but not /t/ and /t̪/ (a dental /t/ made with the tongue at the teeth). Japanese learning infants do not need to discriminate either contrast, and Hindi learning infants must eventually discriminate both. A wealth of research suggests this problem is solved early. By 6 months, infants can discriminate many of the speech contrasts used across the world's languages (Kuhl, 1979; Werker & Tees, 1984). By roughly the first birthday, this ability narrows to encompass mostly native-language contrasts (Kuhl, Stevens, Deguchi, Kiritani, & Iverson, 2006; Tsuji & Cristia, 2014; Werker & Polka, 1993; though see Best, McRoberts, & Sithole, 1988; Eilers & Minifie, 1975; Narayan, Werker, & Beddor, 2010), and the ability to discriminate native contrasts is enhanced (Galle & McMurray, 2014; Kuhl et al., 2006; Tsuji & Cristia, 2014). There are additional changes in the second year (Dietrich, Swingley, & Werker, 2007; Hay, Graf Estes, Wang, & Saffran, 2015; Rost & McMurray, 2010), as children learn which acoustic dimensions are *relevant* for language. These abilities—what Hay et al. (2015) term "interpretive narrowing"— converge on the native language by 24 months. This body of work has led to the canonical view that speech categories are stable and native-like by 24 months.

The early development of these skills constrains the mechanisms that might underlie this development. During infancy, there are neither robust speech production abilities, nor many words in the lexicon to shape perceptual development. This argues for some form of perceptual or statistical learning, based on the acoustic signal alone (Jusczyk, 1993; Kuhl et

al., 2005; Werker & Curtin, 2005). The most prominent theories (de Boer & Kuhl, 2003; Maye, Werker, & Gerken, 2003; McMurray, Aslin, & Toscano, 2009) suggest that infants attend to the statistical distribution of phonetic cues, and use the clustering of cue values to identify categories.

Speech perception in older children

Several studies of preschool- and school-age children challenge the view that speech perception develops early (Bernstein, 1983; Hazan & Barrett, 2000; Nittrouer, 1992, 2002; Nittrouer & Miller, 1997; Nittrouer & Studdert-Kennedy, 1987; Slawinski & Fitzgerald, 1998). These studies typically use paradigms in which children label tokens from a continuum spanning two phonemes. For example, the most important cue for categorizing a sound as voiced (e.g., /b, d, g/) or voiceless (e.g., /p, t, k/) is Voice Onset Time (VOT) which reflects the time difference between release of the lips or tongue and the onset of laryngeal voicing. In English, voiced sounds like /b, d, g/ show low VOTs near 0 msec, and voiceless sounds are indicated by VOTs near 50 msec. Typical studies with older children manipulate VOT in small steps and ask listeners to categorize each token as /b/ or /p/ in order to precisely characterize categorization.

Many studies using this approach examine how children's ability to weight and combine different cues develop (e.g., use of VOT and secondary cues like pitch; Bernstein, 1983; Nittrouer & Studdert-Kennedy, 1987). These suggest an extended period of learning which cues are important. However, even studies focusing on single cues (Hazan & Barrett, 2000; Slawinski & Fitzgerald, 1998) show steeper identification functions over development (Figure 1). These studies vary in when they see developmental change, with some revealing change up to age six (Nittrouer, 2002; Nittrouer & Studdert-Kennedy, 1987; Slawinski & Fitzgerald, 1998), but with others showing later changes (Hazan & Barrett, 2000; Nittrouer, 2004).

Continued development after infancy raises the possibility that other developmental mechanisms (beyond bottom-up learning) drive speech development. For example, a growing lexicon (Feldman, Griffiths, Goldwater, & Morgan, 2013; Metsala & Walley, 1998) or phoneme awareness training during reading instruction (Dich & Cohn, 2013) could both alter phoneme categorization. To build such a theoretical account, however, we must answer a critical question: what exactly is changing with development? That is, when we observe a steeper categorization slope over development, what changes in perception and language are responsible?

The canonical view is that shallow identification slopes in younger listeners reflect more gradient (less discrete) representations of categories. Increasingly categorical responding with age is consistent with classic thinking that discrete categories support better perception. However, this creates a puzzle. The current standard view in adult speech perception is that speech categories are gradient (Andruski et al., 1994; McMurray et al., 2002; Miller, 1997) and this makes perception more flexible (Kapnoula, Winn, Kong, Edwards, & McMurray, 2017; McMurray, Tanenhaus, & Aslin, 2009). In this light, children appear to develop toward a discrete mode of categorization that is neither observed nor ideal in adults.

Categorical or Gradient Speech Perception?

Work using goodness ratings suggests adult speech categories reflect a graded prototype structure (Miller, 1997) with some cue values (e.g., specific VOTs) being better exemplars of a category than others. Such gradiency is even observed in infants (Galle & McMurray, 2014; McMurray & Aslin, 2005; Miller & Eimas, 1996). Gradiency likely derives from developmental history, reflecting the gradient statistical distributions of speech cues (Clayards, Tanenhaus, Aslin, & Jacobs, 2008; McMurray & Farris-Trimble, 2012; Miller & Volaitis, 1989).

Gradient categories may play a number of functional roles (Kapnoula et al., 2017). They could help listeners calibrate the degree of commitment to a phoneme category to the likelihood that that phoneme is the correct interpretation: when the input is near prototypical values listeners should commit fully, but when it is more ambiguous it may be helpful to be more cautious. This helps listeners "hedge" their bets, keeping options open when the input is ambiguous in case they need to revise an earlier decision (McMurray, Tanenhaus, et al., 2009). Gradiency also implies a more fine-grained analysis of within-category details, enabling listeners to get more out of the signal, such as coarticulatory information that can help them anticipate future sounds (Gow, 2001). Indeed, Kapnoula et al. (2017) recently showed that listeners who are more gradient are also better at integrating secondary cues.

If listeners should be striving for more gradient representations, what should we make of steeper (less gradient) categorization slopes with development? One possibility is that the slope of categorization function may not reflect the nature of the categories used in perception (Supplement S1 for a model). A shallower slope could also reflect ***noisier encoding of acoustic cues*** like VOT, even if the structure of the category (the way cue values are mapped to categories) was discrete. For example, if a VOT of 18 msec (a /b/, but near the boundary of /b/ and /p/) is occasionally encoded as 24 msec (/p/), this would result in more /p/ responses. However, if a 0 msec VOT was miscoded as 8, this would not change the category (both are /b/'s). Thus, noise in cue encoding could flatten the function near the boundary. Development may then derive from reductions in encoding noise, not changes in category structure.

The converse is also true. Even if listeners had a gradient mapping (e.g., their mapping reflected the fact that a 0 msec VOT is 100% /b/, but an 18 msec is 60%), they must map this gradient phoneme activation to the responses in the experiment. Older listeners may choose the highest probability category on every trial ("winner take alll"; which may be optimal: Nearey & Hogan, 1986), while younger listeners might probability match, choosing different options for that same VOT from trial to trial to reflect this uncertainty. Under this view, developmental change is outside of speech perception, but in the ***response system*** (perhaps part of a cognitive control system) that maps phonological activation to experimental responding.

The point of both of these examples that a shallower or steeper phoneme categorization slope in a 2AFC task may not entirely reflect the structure of the child's phoneme categories. What is needed is a way to isolate gradiency in the underlying mappings from cues to categories. One source of evidence for gradiency in adult speech comes from work

using eye-tracking in the visual world paradigm (VWP; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995) to examine speech categorization (McMurray, Aslin, Tanenhaus, Spivey, & Subik, 2008; McMurray et al., 2002). This may offer ways to isolate category structure.

McMurray et al. (2002) presented adults with speech continua spanning two words (e.g., *bear*/*pear*), who selected the referent of the word they heard from a screen containing the endpoints of the continuum and unrelated items. The overt decisions are analogous to standard phoneme identification decisions, and showed the typically steep slope. While subjects performed this task, eye-movements to the referents were monitored. Eye-movements are generated as early as 200 msec after word onset, with multiple looks over the trial, reflecting listeners' unfolding commitment to possible interpretations. As VOT approached the boundary, participants made increasing looks to the competitor (Figure 2A for a schematic). This was true even in a conservative analytic approach that examined only trials in which the listener responded "correctly" for a given stimulus (relative to their personal boundary). Thus, evidence of gradiency could be seen even though all trials in the analysis were categorized identically.

This version of the VWP may isolate developmental changes in the gradiency of the underlying cue→category mapping (the structure of the category) from differences in cue encoding or the response system. Under a noisy cue encoding approach, a shallow identification slope comes from the trials in which the VOT was 18 msec (/b/) but misheard as 24 msec (/p/). However, those trials would receive the incorrect response and be excluded from this analysis. Similarly, if changes in the slope derive from how graded phoneme activation is linked to the response, eye-tracking may reveal the true structure of the category, since it is more implicit. Thus, if developmental changes derive solely from noisy cue encoding or differences in decision level processes, when we minimize their effects with this approach there, should be few developmental differences in gradiency. Conversely, if the underlying mappings are developing, we may observe that children become more or less gradient with development.

Lexical Processes

This application of the VWP raises an additional potential locus for development: lexical development. Lexical and perceptual processes are difficult to divorce: work in adults suggests preliminary states of categorization cascade immediately to lexical access (Andruski et al., 1994; McMurray et al., 2002; Utman, Blumstein, & Burton, 2000). Thus, lexical development may contribute to what looks like development of speech categorization. In McMurray et al. (2002), and many studies with children (e.g., Hazan & Barrett, 2000; Nittrouer, 2004; Slawinski & Fitzgerald, 1998), this is even more likely as the stimuli are familiar words (e.g., *beach/peach*).

Lexical access is thought to be based on competition (Dahan & Magnuson, 2006; Weber & Scharenborg, 2012). As the input arrives, multiple words that partially match the signal are briefly active (e.g., as listeners hears *peach,* they activate *peak, beach,* and *peel* [etc.]). These words compete as the input unfolds, until one remains. A wealth of evidence documents growth in the efficiency of lexical competition during late infancy (Fernald, Perfors, &

Marchman, 2006; Law, Mahr, Schneeberg, & Edwards, 2017; Zangl, Klarman, Thal, Fernald, & Bates, 2005). More relevant, Rigler et al. (2015) showed changes between 9 and 16 years of age (and see Sekerina & Brooks, 2007). They show increasing efficiency of activating the target, and faster suppression of competitor with development. Given the close links between speech perception and word recognition, lexical development may explain some changes in speech perception. In particular, *lexical inhibition* (Dahan, Magnuson, Tanenhaus, & Hogan, 2001) – by which more active words suppress less active competitors could act in much the same way as the response system described above to "clean up" more ambiguous phoneme level activation.

Present Study

This study used the McMurray et al. (2002) paradigm to test three age groups in a cross-sectional design: 7–8, 12–13 and 17–18 year olds. Children heard tokens from several speech continua (e.g., *bear/pear*) and selected the referent from a screen containing both endpoints, and an unrelated minimal pair (e.g., *sip/ship*). Eye-movements were monitored to measure of how strongly each word was considered over time. In addition to voicing (VOT), we also examined fricative place of articulation (*s/ʃ*), as several studies document changes in fricative perception during these years (Hazan & Barrett, 2000; Nittrouer, 2002; Nittrouer & Miller, 1997).

A few studies have used this paradigm to examine individual (but not developmental) differences (McMurray, Farris-Trimble, Seedorff, & Rigler, 2016; McMurray, Munson, & Tomblin, 2014). These studies generally interpret the slope of competitor functions (Figure 2) as a marker of the structure of the mapping between cues and categories, and the overall height as a marker of the degree of lexical competition over and above differences in speech perception. As in prior work, we analyzed eye-movements relative to each child's own boundary and to their response, to minimize the role of differences in encoding noise and the response system on this assessment of category structure and lexical competition. We addressed two critical questions:

1. *What is the developmental time course of the sharpening of phonetic categories?* We examined the mouse clicking responses similar to standard phoneme decision tasks. As in prior studies (Hazan & Barrett, 2000), we expected steeper categorization slopes with development (Figure 1), though we tested a larger age-range. Importantly, we extended this by constructing an analogue of identification curves from the fixation record to capture categorization as it unfolds over milliseconds *before* the response. This more sensitive measure of categorization slope may reveal later development than seen previously. As real-world language unfolds at a high rate, the preliminary states of processing tapped by this measure may provide a better view of the skills that children actually bring to bear during real language processing.
2. *Does the development of phoneme categorization derive from changes in the way continuous inputs are mapped to speech categories, and what is the form of those changes?* We next examined looking relative to each child's own category boundary to identify how the structure of the phonetic categories change with

development. There are four possibilities. First, if the major determinant of identification slope is noise in the encoding cues, once this is accounted for, we would observe no developmental differences, but an effect of VOT. Second, changes in categorization slope with development may mirror changes in the category structure, with children becoming less gradient in both measures with age (Figure 2B). Given that adults are gradient (Andruski et al., 1994; McMurray et al., 2002), older children are not likely to be fully categorical, but there is room for even more gradiency in young children. Third, the opposite pattern is possible: younger children may show less sensitivity to within-category detail, and become more gradient with age (Figure 2C). This would seem to conflict with a shallower categorization slope in young children (though this could derive from noise or response system changes). In this case, increasing sensitivity to fine-grained differences in speech cues may ultimately enable a sharper and more confident end-state decision. This is supported by a recent meta-analysis of studies of VOT discrimination in infancy which found a steady increase in *within-category* discrimination over the first year (Galle & McMurray, 2014). Finally, we also expected to see changes in lexical competition (Rigler et al., 2015) with heightened competition for younger listeners. Absent changes in category structure, this would appear as the pattern in Figure 2D, where the effect of VOT (or frication step) is the same in younger and older participants, but younger participants show more activation overall for competitors. This hypothesis is not mutually exclusive of the other hypotheses.

Methods

Participants

Seventy-six children participated in this experiment. We tested three age-groups: 7–8 year olds (N=25), 12–13 year olds (N=26), and 17–18 year olds (N=25). Two in the middle group were excluded for difficulties with the eye-tracker. Children were recruited using mass e-mails to the University of Iowa community and the University of Iowa Hospital newsletter, in accordance with the university human subject protocols (University of Iowa IRB# 201207756, *Language Processing in Adolescents and Children*). Participants were compensated $15/hour.

Sample size was determined by computing minimum detectable effects for a given sample. While there is no closed solution for power in mixed models (the intended statistical approach) we approximated it with a mixed ANOVA (widely assumed to be less sensitive) with age group (3 levels, between) and VOT (4 levels, as in the final analyses reported here, within). Assuming α=.05, β=.8, ρ=.5, and N=25/group, this led to an MDE of d=.296 for the effect of VOT, d=.579 for the age effect, and d=.301 for an interaction).

Participants reported normal hearing and were native monolingual English speakers. All children were typically developing with normal or corrected-to-normal vision by parent report. We ran a small battery of language and nonverbal assessments to quantify individual differences. A hearing screening was conducted on all participants. Sixty-eight participants passed the hearing screening with better than 25 dB hearing at four frequencies (500,

1000500, 2000, 4000 Hz). Six participants failed only at the 500 Hz frequency (all at 30 dB). As this was below the threshold for clinical intervention, they were retained. One participant failed at all frequencies and all degrees of loudness, but was able to communicate naturally, indicating a lack of understanding of the audiogram; he was retained.

Standardized Assessments—We assessed receptive vocabulary with the *Peabody Picture Vocabulary Test* (PPVT-IV; Dunn & Dunn, 2007). This was not available for one participant (in the 7–8 y.o. group). To measure overall language, we administered two subtests (*Recalling Sentences* and *Understanding Spoken Paragraphs*) of the *Clinical Evaluation of Language Fundamentals* (CELF-4; Semel, Wiig, & Secord, 2006). The *Understanding Spoken Paragraphs* subtest was not administered to the 7–8 y.o.s because this it is not normed for these ages. CELF scores were not available for 5 participants (all in the 7–8 y.o. group). To assess nonverbal IQ, we administered the *Block Design* and *Matrix Reasoning* subtests of the *Wechsler Abbreviated Scale of Intelligence* (WASI-II; Wechsler & Hsiao-pin, 2011). Matrix reasoning scores were not available for 4 participants (in the 7–8 y.o. group) and Block Design scores were not available for 2 participants (in the 7–8 y.o. group) because of time constraints.

All three age groups were within the normal range (Table 1), and no child scored below a standard score of 90 on the PPVT or 82.5 on the CELF. We conducted one-way ANOVAs to assess differences among the age-groups in standard scores, with follow-up tests comparing adjacent ages. We found a significant effect of age on PPVT ($F(2,72)=5.9$, $p=.004$) with differences between 12–13 and 17–18 y.o. ($t(70)=2.3$, $p=.022$), but not between the 7–8 and 12–13 ($t<1$). We did not find a significant effect for CELF ($F(2,68)=2.0$, $p=.144$). However, there was a significant effect for WASI ($F(2,71)=3.6$, $p=.032$) with 7–8 y.o. performing higher than the 12–13 y.o. ($t(69)=2.64$, p-=.01), but no difference between the older groups ($t<1$). Overall, scores were average or above and group differences were not in the same direction across ages, nor seen in all measures. Thus, groups were reasonably balanced. To guard against spurious group differences, however, we included these scores as covariates in our analysis.

Design

This experiment used six minimal pairs differing in fricative place of articulation (/s/ vs. /ʃ/: *shack/sack*, *shave/save*, *self/shelf*, *sign/shine*, *sip/ship*, and *sock/shock*) and six differing in voicing (/b/ vs. /p/: *beach/peach*, *bear/pear*, *bet/pet, bin/pin, bug/pug*, and *bump/pump*). Minimal pairs were real words that were easily portrayed in pictures and likely known by the youngest age-group. For each pair, an eight-step continuum was constructed. On each trial, participants heard one token and selected the referent from a screen containing both endpoints of the target continuum, and those of a continuum from the other class. Thus, a fricative pair served as unrelated foils on voiced trials, and vice versa. To avoid emphasizing the relationship between members of a minimal pair, the pairing of specific voicing and fricative pairs was fixed throughout the experiment and randomly selected for each participant. Each continuum step was heard six times, resulting 2 (continua types) × 6 (continua) × 8 steps × 6 reps = 576 trials.

Stimuli

Auditory Stimuli were constructed from natural recordings using techniques similar to McMurray et al. (2016). These were based on recordings of a male native English speaker with a standard Midwest dialect, recorded in a sound attenuated room with a Kay CSL 4300B at 44.1 kHz. Multiple exemplars of each word were recorded in a carrier phrase (*He said X*). Target words were then isolated from the phrase to construct the continua.

VOT continua were created using progressive cross-splicing (McMurray et al., 2008). One exemplar of each endpoint was selected that best matched on pitch, duration and formant frequencies. Next, a predetermined duration was deleted from the onset of the voiced token (e.g., *beach*) and replaced with the corresponding segment from the voiceless token (*peach*). This was done at approximately 8 millisecond increments, with splice points at the closest zero-crossings. This led to 8-step continua ranging from 0 to 56 msec of VOT.

Fricative continua were created using a spectrum shifting technique developed by (Galle, 2014; McMurray et al., 2016; Supplement S2 for a more thorough description). This was done by first extracting the frication from the original recordings, and computing the long-term average spectra of those segments. Spectra were then shifted (in frequency space) in 8 equal steps, and the final fricatives were constructed by filtering white noise through these shifted spectra. Code is available at https://osf.io/vz6wp/.

Visual Stimuli were developed using a standard laboratory procedure to ensure representative images of the auditory stimuli. For each word, several images were selected from a clipart database and a focus group of graduate and undergraduate students determined the most prototypical image. These images were then edited to minimize visual distractions, use more prototypical colors or other features, and to ensure a uniform style. Each image was approved by one of three members of the laboratory with extensive experience using the VWP.

Procedure

Stimuli were presented over Bose loudspeakers amplified by a Sony STR-DE197 amplifier. Volume was initially set at 70 dB, and participants adjusted it to a comfortable level during a brief training. A padded chin rest, 29″ from the screen, minimized head movements.

After obtaining informed consent, participants were seated in front of a 1280 × 1024 17″ computer monitor. The experimenter adjusted the chin rest to a comfortable position. Children were told that they could relax during breaks (every 32 trials), but were to return to the chin rest during testing. The researcher then calibrated the eye-tracker, and verbal instructions were given. Before testing, participants completed an eight-trial training to become familiar with the task and adjust the volume if needed. Auditory and visual stimuli in training differed from those in testing. Participants then completed a second phase of training to familiarize them with the visual stimuli. During this phase, participants advanced through each of the 24 images paired with the written word and a natural auditory recording.

On each testing trial, participants saw four pictures (one in each corner) with a small red circle in the center of the screen. Images were 300×300 pixels, 50 pixels from the edge of

the screen. Trials started with a 500 msec preview so participants were aware which pictures were present on that trial and their locations, minimizing the role of visual search on subsequent eye movements. After 500 msec, the circle turned blue and the participant clicked on it to hear the stimulus. Subjects then clicked on the corresponding image using the mouse. Participants were encouraged to take their time and perform the task naturally.

Eye-Movement Recording and Analysis

Eye-movements were recorded using a SR Research Eyelink 1000 desktop mounted eye-tracker. Eye-movements were recorded at 500 hz and down-sampled to 250 hz for analysis. The standard 9-point calibration was used. A drift correction procedure was conducted every 32 trials to account for the natural drift in the eye-track. If the participant failed a drift correction, the eye-tracker was immediately recalibrated. Eye movements were automatically classified into saccades, fixations, and blinks using the default Eyelink parameters. Events were grouped into “looks” which began at the onset of the saccade and ended at the end of the subsequent fixation (McMurray et al., 2008). Trials were variable length (ending when the participant made a response). To cope with this, the eye-movement record was fixed to 3000 msec; for trials ending before then, the last fixation was extended to 3000 msec; for trials ending after 3000 msec, the fixation record was truncated. Image boundaries were extended by 100 pixels to account for noise in the eye-track. This did not result in any overlap in the regions of interest.

Results[1]

The first analysis examined mouse-click data, analogous to phoneme identification tasks. The second developed an analogue of identification data from the eye-movements to examine how identification unfolds over time. These analyses address Question 1, the developmental time-period over which speech categorization sharpens. The third analysis examined looking as a function of the continuum step controlling for the participant’s own category boundary and their response on each trial. This identified developmental changes in the gradient mapping between cues and categories (Question 2). A final analysis reported in Supplement S4 examines the detailed timecourse of lexical competition. All analyses investigated standardized test scores as moderators. Given missing data and the collinearity between measures, the four language measures (two CELF subtests, PPVT and EVT) were averaged into a composite language score. For one participant with no language scores, we used the mean of her age group (7–8 y.o).

Identification

Stop Voicing—Trials in which the participant selected a non-b/p word were eliminated (7–8 y.o.: 30 trials, M=1.30 trials/participant; 12–13 y.o.: 30 trials; M= 1.25 trials/participant; 17–18 y.o.: 6 trials, M= 0.24 trials/participant). Figure 3A shows the proportion /p/

[1]Initially this study collected data on 58 children. Data were analyzed and submitted for publication. Between study design and submission however, many top-tier journals began requiring a minimum sample size. Thus, we initiated a second round of additional data collection in Spring, 2017, which is reported here. As this violates best practices in null hypothesis testing, for scientific and statistical openness, the original methods and results are posted in a public repository (https://osf.io/mqeh4/). There are no differences in the conclusions.

responses by step and age-group. All three age groups successfully categorized the stimuli, with asymptotic performance near 0 (for low VOTs) and 1 (for high VOTs) and a boundary near step 3 (~15 msec of VOT). There was a small shift in the slope as well, with shallower slopes for younger children. With proportional data, however, differences in the slope of the average could derive from variability in boundaries among individuals. If each subject had a steep slope, but younger listeners showed more boundary variability, the group slope could appear shallower. At analysis, this was handled with logistic regression, which accounts for this on a subject-by-subject basis. For visualization, we recomputed the identification results as a function distance from each participants' own estimated boundary using a procedure described below (relative step or rStep). Figure 3B shows this small shift in slope is still present in this more conservative analysis.

These data were analyzed with a binomial mixed effects model using lme4 (ver 1.1–12) in R (ver 3.3.1) (Supplement S3 for an analysis which directly estimates slopes using curvefitting). Fixed effects included step (centered), two contrast codes for age-group, and the step × age-group interaction(s). The first age code contrasted young (7–8 y.o.) and middle (12–13 y.o.; −1/+1, old=0, centered). The second contrasted middle and old (17–18 y.o.; −1/+1, young=0, centered). Language (centered) was also included. Potential random effects included participant and item. We compared random effect structures using AIC to find the best fit for the data. The optimal random effects structure included a random intercepts of subject and item, and a random slope of step on subject and word. We dropped covariance terms between the intercept and slope on subject to avoid non-convergence. The final formula is shown in (1)

$$\text{P}\sim\text{Step} * (\text{young v middle} + \text{middle v old} + \text{Language}) + (1 \mid \text{Subject}) + (0 + \text{Step} \mid \text{Subject}) + (1 \mid \text{Item}) + (0 + \text{Step} \mid \text{Item}) \quad (1)$$

Results are shown in Table 2 (top). This model showed a significant effect of step ($p<.0001$). There were also main effects of both age contrasts ($p<.0001$). This was due to small shifts in the category boundary across the ages. Crucially, there were significant interactions between step and young vs. middle ($p<.0001$) and middle vs. old ($p<.0001$). This indicates that the categorization boundary was steeper for the middle group than the young group, and that the sharpening continued between the middle and old groups. The effect of step interacted with language ($p=.01$) indicating a somewhat steeper slope for children with better language.

Fricative Place of Articulation—Trials were removed if the subject selected a non-s/ʃ item (young: 43 trials; M=1.87 trials/participant; middle: 25 trials, M=1.04; old: 6 trials; M=0.24). Fricative identification showed the same pattern as voicing (Figure 3C, D): listeners performed well on the endpoints with a steep transition and small changes in slope with age. We used the ame statistical model as for voicing (Equation 1). There was a significant effect of step ($p<.0001$). There were no main effects of either age contrast,

suggesting similar boundaries at all three ages. We found a significant step × age interaction for the young vs. middle contrast (p=0.014) and a marginal interaction for middle vs. old (p=.07). This supports a steepening of the slope with age. A significant interaction between step and language (p=0.0075) indicated that individuals with better language tended to have steeper slopes.

Summary—We found a steeper identification slope between 7 and 12 for both voicing and fricative continuums. Moreover, the older age group showed continued sharpening for voicing, with marginal evidence for fricatives. Both continua also showed some moderation by individual ability with sharper slopes for listeners with better language abilities.

Overview of Eye-Movement Analysis

Analysis of the eye-movements started by computing the proportion of trials on which the participant was looking at each competitor at each 4 msec time slice for each condition. Figure 4 shows timecourse functions for endpoint stimuli. We designate the object consistent with the current step as the *target* (e.g., for a 0 msec VOT, the /b/ item), and its minimal pair the *competitor* (e.g., the /p/ item). Meaningful fixations began at about 300 msec, reflecting 200 msec to plan an eye-movement, plus 100 msec of silence at stimulus onset. After that, looks to the target increase, and competitors receive more looks than unrelated objects briefly. Fricatives (Panels D–F) showed more competition, and a slower timecourse than stops (Panels A–C), consistent with prior studies (Galle, 2014; Galle, Klein-Packard, Schreiber, & McMurray, submitted). A few developmental patterns are apparent (Supplement S4 for complete analyses). First, as in Rigler et al. (2015), the slope of target looking increases with age. Second, competitor looks (relative to unrelated items) increase slightly with development. For stops, at 7–8, competitors are barely fixated more than the unrelated, whereas by 12–18 years they receive more looks. Fricatives may be more complex with changes in both the degree and timing.

The temporal unfolding of speech categorization and its development

Mouse click results suggested speech categorization continues to develop through adolescence, though effects were small. We extend that by using the fixation record to ask how this categorization unfolds in real-time—between the moment of hearing the stimulus, and the response—and how these dynamics change with development. This may offer a more sensitive way to address how late speech categorization develops (Question 1), as this analysis may detect developmental differences that affect early processing, but not late processing.

We first computed how strongly participants were committed to one candidate at each moment during processing. For voicing, $bias_{bp}$ was the average looks (across trials) to the /p/ items minus those to the /b/ items. For fricatives, $bias_{\int s}$ was looks to the /s/ items minus those to the /ʃ/ items. Bias was computed each 20 msec for each continuum step (ignoring the response).

Figure 5 shows a visualization of this (for cool animations, see http://osf.io/w5bqg). By the end of processing, all groups showed steep categorization, consistent with mouse clicking

(Figure 3). But there were developmental differences. For voicing (Figure 5, A–C), by the end of processing (~1400 msec), there were differences in the asymptotes between 7–8 y.o. and 12–13 y.o.. There are also differences at the intermediate time points. At 600 msec (Figure 6A), 7–8 y.o,s are barely off chance, while 12–13 and 17–18 y.o.s have made a partial commitment. At 800 msec, 7–8 y.o, are far from their ultimate level of performance, while 12–13 y.o. are much closer, and 17–18 y.o. are indistinguishable from it. For fricatives, processing is delayed by several hundred milliseconds and with marked developmental differences[2]. These are seen at 800, 1000, and 1200 msec, where even the 12–13 y.o.s and 17–18 y.o.s differ (Figure 6B).

For analysis, we fit a four- parameter logistic (2) to the data for each participant at each time (e.g., Figure 6), to estimate the shape of the categorization function at that time.

$$P(target) = \frac{p - b}{1 + \exp\left(4 \cdot \frac{s}{p - b} \cdot (c - step)\right)} + b \quad (2)$$

This equation describes a sigmoidal function of continuum *step* that starts at a lower asymptote (baseline or *b*), and transitions to an upper asymptote (or peak, *p*). The crossover is described by *c*, and the slope at the crossover by *s.* This function was fit to each participants' bias at each time using a constrained gradient descent algorithm (McMurray, 2017). Fits were good (b/p: average r = .926; ʃ/s: r = .892). We excluded a small number of fits (b/p: 138/7474; ʃ/s: 274/6816) with poor correlations (r<.4). These were almost entirely in the first few hundred milliseconds when the data were noisy, and were evenly distributed across age groups.

Our analysis focused on two parameters of the logistic that describe categorization. First, we asked how the *categorization slope* (*s*) changes with time and age. Second, we examined the separation of the asymptotes or *categorization amplitude* (*p* – *b*). This offers a measure of the confidence of the decision. Figure 7 shows the categorization slope and amplitude over time and age. Developmental effects were seen for both continua. The largest differences are between 7–8 and 12–13 y.o.s where even at the end of processing (~2000 msec) differences are seen for both continua. However, particularly for amplitude, there were earlier points in processing (e.g., prior to 1300 msec) where 12–13 y.o.s differed from 17–18 y.o.s. Fricatives (Panels B, D) show a much shallower categorization slope overall than voicing, and are delayed to reach peak.

We averaged slope or amplitude across 100 msec to examine a smaller number of times. We then compared slope or amplitude at these time-points between adjacent ages (Table 3). Because we were making a large number of comparisons (but did not want to make strong claims about any one), p-values were adjusted to maintain a constant False Discovery Rate (Benjamini & Hochberg, 1985), rather than with a conservative family-wise error correction.

[2]This may derive at least in part from the fact that fricatives occur less frequently in children's input than stop consonants, with concomitant delays in development at earlier ages (Thiessen & Pavlik, 2016).

For voicing, the youngest group differed from the middle group in both slope and amplitude throughout the timecourse of processing, and the middle and older group differed in amplitude through approximately 700 msec. This can be seen in Figure 6A which shows $bias_{bp}$ at each age group at 600 msec. Similarly, for fricative place we observed differences between the younger and middle group at all times, but only marginally significant differences between the middle and older group at the intermediate times points (i.e. 700–1000, Figure 6B).

For a deeper analysis of the real-time changes in categorization, we further fit a four parameter logistic to the categorization slope and amplitude estimates over time. This was done separately for the b/p and ʃ/s continua for each participant. The resulting parameters were then compared between adjacent age groups to determine the effect of age on the timecourse of the slope and amplitude of the categorization function (Table 4).

For b/p, categorization amplitude (Figure 7A) grew slower over time (had a later crossover and a shallower slope) in 7–8 than 12–13, and in 12–13 than 17–18 y.o. children. It also reached a lower ultimate value in the 7–8 vs. the 12–13 y.o. age groups, but did not differ for the older groups. We found fewer differences on the slope of the b/p categorization function (Figure 7C); however, the ultimate slope (asymptote) was lower in 7–8 than 12–13 y.o. children. The ʃ/s continuum showed similar results. Categorization amplitude (Figure 7B) grew slower in 7–8 y.o. and 12–13 y.o. children, and between 12–13 and 17–18 y.o. children. Amplitude also had a marginally lower ultimate asymptote in 7–8 y.o. than 12–13 y.o. (no difference between older groups). Categorization slope in fricatives (Figure 7D) showed differences between the 7–8 and 12–13 age-groups in the ultimate asymptotic slope, but no other differences.

Together, these analyses suggest that the dynamics of speech categorization differ well into adolescence. While the most robust differences were seen in the dynamics of the categorization amplitude over time, younger listeners also differed in slope.

Development of Within-Category Sensitivity

Our final analysis investigated the structure of children's phonetic categories over and above changes in overt identification. As described, a shallower identification slope (in younger children) could arise even with no difference in the mapping between cues and categories, for example, if noise in VOT encoding causes tokens near the boundary to be miscategorized. To more directly assess the gradiency of the mappings, we used a technique from prior work (McMurray et al., 2008; McMurray et al., 2014; McMurray et al., 2002) that examines sensitivity to VOT (or other cues) while controlling for the response. This eliminates trials in which encoding noise caused the subject to make the wrong response.

For this, we first recoded continuum step as distance from each participant's boundary (Relative Step [rStep]). This eliminates the possibility that variability in the boundary across subjects could make an age-group look more gradient. To compute boundaries, we fit the four-parameter logistic (2) to each child's identification data, separately for b/p and ʃ/s continua. Since the boundary also varies across words, we fit logistic functions to each item (averaged across participants). Then for each trial, we computed the boundary by adding the

participant's boundary to the deviation of the item boundary from the grand mean. Continuum step was then recomputed as rStep, the distance from that boundary. Negative rSteps corresponded to /b/ and /ʃ/, and positive to /p/ and /s/. Finally, we removed trials on which the participant chose the competitor (e.g., for negative rSteps, /p / or /s/; for positive rSteps, /b/ or /ʃ/). Finally, we examined looks to the competitor to estimate how strongly competitor looking was sensitive to within-category VOT (or frication step) differences with VOT or frication step now coded as distance from that participants own boundary (rStep).

This procedure isolated only trials in which tokens were clearly on the target side of the continuum and the target was chosen to somewhat isolate differences in category structure from other factors. Figure 8 shows looks to the competitor as a function of time and rStep, averaged across all ages. As in prior studies, we see a gradient response: for rSteps far from the boundary (±3) participants made few competitor looks, and this increased as rStep approached the boundary (moved toward 0). For statistical analyses, we computed the area under the curve (AUC) at each rStep. We averaged over 300–2300 msec for stops and 600–2600 msec for fricatives. The 300 msec onset for b/p stimuli is consistent with prior work (McMurray et al., 2002) and reflects 100 msec of silence at stimulus onset, and a 200 msec oculomotor delay; the later onset for fricatives is based on work suggesting that lexical access does not begin until the offset of fricatives (Galle et al., submitted; McMurray et al., 2016).

Figure 9 shows AUC as a function of rStep and age. It suggests a large difference in overall competitor looks, with younger children showing more competitor activation than 12–13 and 17–18 y.o.s. It also suggests differences in the degree of sensitivity to within-category differences, with younger children showing very little effect of rStep and the older age groups showing more. As heightened within category sensitivity is thought to mirror more gradient categories (McMurray et al., 2002), this would appear to conflict with the shallower slope of the mouse-click identification functions (which we return to in the discussion).

Statistical analyses used these AUC estimates as the dependent variable in a linear mixed effects model. Separate AUC estimates were computed for each participant, for each continuum (item), at each rStep. AUC was log-transformed to eliminate skewed residuals. Each model included fixed effects of rStep (centered), and a quadratic term ($rStep^2$) to account for non-linearities in the response to rStep. Age group was a between-subject fixed effect (coded, as before, in terms of two contrast codes), and we assessed the age group × rStep interaction. As in prior work (McMurray et al., 2016), we included looks to the competitor when it was unrelated (e.g., looks to the /b/ object when the stimulus was /s/ or /ʃ/) as a covariate (centered) to account for differences between participants and items in overall looking.

Random effects were based on prior studies (McMurray et al., 2016; McMurray et al., 2014) using this paradigm. These included random intercepts for subject and item, as well as random linear and quadratic slopes of rStep on subject. Because not all subjects had the same number of eligible trials, we weighted the model responses based on the number of trials contributing to the value. The resulting model in shown in Equation 3.

$$\text{AUC}\sim \text{YvM} * (\text{rStep} + \text{rStep}^2) + \text{MvO} * (\text{rStep} + \text{rStep}^2) + \text{Language} * (\text{rStep} + \text{rStep}^2) + \text{Unrelated} + (\text{rStep} + \text{rStep}^2 \mid \text{subject}) + (1 \mid \text{item}) \quad (3)$$

Separate versions of this model were used for each type of continuum and for each side.

Stop Voicing—We started by examining the voiced side of the VOT continua (Figure 9A, left; Table 5, top). We found a significant effect of rStep ($p<.0001$) and rStep^2 ($p=.0175$) with more looks for tokens near the boundary. We also saw more overall competitor looking in 12–13 than 7–8 y.o.s ($p=.0001$), and in 17–18 than 12–13 y.o.s ($p=.0036$). Finally, there was a nearly significant interaction between the linear effect of rStep and the young vs. middle contrast ($p=.052$) indicating reduced sensitivity to VOT in the 7–8 than 12–13 year olds.

We next examined the voiceless side of the continuum (Figure 9A, right; Table 5, bottom). Again, we found significant effects of rStep ($p<.0001$), and a quadratic effect ($p=.00025$), accounting for the flattening of the function at high rSteps. We also found increased overall competitor looking for younger subjects at both age contrasts (YvM: $p<.0001$; MvO: $p<.0001$). Children with poorer language showed increased competitor looks ($p=.0083$), though this did not interact with rStep. There were significant interactions of rStep with both age contrasts (YvM: $p=.042$; MvO: $p=.046$), and marginal interactions with the quadratic effect of rStep. With age, participants showed increasing sensitivity to gradient differences in VOT.

Fricative Place of Articulation—We next examined the /ʃ/ side of the fricative continua (Figure 9B, left; Table 6, top). We found significant effects of rStep ($p<.0001$), and rStep^2 ($p=.0008$) indicating increasing competitor looks as rStep approached the boundary. We also found significant differences in overall looking at both age contrasts (YvM: $p=.01$; MvO: $.025$). The interaction between rStep and the middle vs. old age-group was significant ($p=.0098$) indicating greater sensitivity to fine-grained detail in the older listeners. Finally, we examined the /s/ side (Figure 9B right; Table 6, bottom). We found significant effects of rStep ($p<.0001$) and rStep^2 ($p=.0025$). There were significant differences in overall competitor looking at both age contrasts ($p<.0001$; $p<.0001$), indicating a general decreases in looks to the competitor with age. The rStep × age was marginally significant in the young vs. middle contrast ($p=.063$).

Summary—These analyses showed a gradient effect of within-category differences on competitor activation, consistent with work on adults (Andruski et al., 1994; McMurray et al., 2008; McMurray et al., 2002) and adolescents (McMurray et al., 2014). We also saw heightened competitor looks at both age contrasts: older children suppress competitors more completely. We also observed (for voicing continua) increased competitor looks in children with poorer language that did not interact with rStep. This replicates (McMurray et al., 2014) showing that language ability affects the degree of competitor looks but does not moderate gradient category structure. Crucially, the degree of gradiency (rStep) interacted with age. For stops, there were differences in gradiency between young and middle age-groups for /b/ and between all three age groups for /p/; fricatives were more limited with

interactions only at the middle vs. old contrast for /ʃ/ and marginally for the younger contrast for /s/. However, there were no interactions in the opposite direction (more categorical with age). Thus, children become *more* sensitive to fine-grained within-category structure over development.

General Discussion

This study had four key findings. First, consistent with prior work (Hazan & Barrett, 2000; Nittrouer, 1992, 2002; Nittrouer & Miller, 1997; Slawinski & Fitzgerald, 1998), the steepness of phoneme categorization (Figure 3) develops throughout childhood. Effects were numerically small, but reliable, and for fricatives, related to language ability. Second, much larger effects were observed in the real-time dynamics of categorization (eye-movements; Figure 5, 6). These were visible throughout the timecourse of processing, but strongest at intermediate times (e.g., 600–1000 msec). Third, we saw development in the degree of lexical competition. This appeared as heightened competitor looks, regardless of VOT or frication step (main effect of age in Figure 9; Supplement S4). Younger children activated competing words more than older children. Finally, over and above that, we found differences in sensitivity to within-category acoustic detail (the slope of the effect of step on looking in Figure 9). We observed *increasing* sensitivity to within-category differences (steeper slope) through 18 years for some conditions. We discuss each of these findings, before turning to potential developmental implications. We start by raising several limitations which may qualify these interpretations.

Limitations

There are four major limitations. First, for both continua, the boundary was not well centered. This was likely due to the fact that stimuli were piloted over headphones, but tested in a soundfield. If the soundfield attenuated low frequencies, this could account for the "left shift" of both boundaries. This is not ideal, but it is unlikely to give rise to our effects. There is no reason why it could lead to spurious changes in slope over development. While the rStep × Age interaction was not observed on each side of the continua, the results did not seem to break down consistently by side. There was a significant effect of rStep for all four analyses, and age × rStep interactions were sometimes only on the short side and sometimes only on the long side.

Second, we only tested two of the many phonetic contrasts relevant for children. We cannot assume that any phonetic contrast is representative of speech as a whole. Categorical perception, for example, is less robust in vowels (Fry, Abramson, Eimas, & Liberman, 1962) and fricatives (Healy & Repp, 1982), and fricatives may be integrated with other portions of the signal differently from other speech sounds (Galle et al., submitted; Ishida, Samuel, & Arai, 2016). Moreover, frequency differences among phonemes could contribute to the rate or robustness of their development (Thiessen & Pavlik, 2016). Thus, these findings should be extended to other speech sounds. However, our primary findings are likely robust. The steepening categorization slope (Figure 3, 5) has been observed for many speech sounds including laterals, vowels, stop consonants and fricatives (Hazan & Barrett, 2000; Nittrouer, 2002; Slawinski & Fitzgerald, 1998). Moreover, within-category gradiency (Figure 9) was

highly similar between stops and fricatives. Thus, both findings are likely robust, though their developmental timecourses may differ across speech sounds.

Third, it is unknown whether the way that lexical representations are mapped to fixations in the VWP changes with development. Such changes cannot be ruled out as a partial source of the developmental changes seen here. However, we note that the one study to systematically examine effects of non-verbal IQ and language on the VPP found no effect of IQ, but significant effects of language (McMurray, Samelson, Lee, & Tomblin, 2010), suggesting some specificity of the link between looking in the VWP and language (though this study only examined adolescents). Future work could address this with purely visual variants of the VWP (Farris-Trimble & McMurray, 2013), or converging language measures like event-related potentials.

Finally, participants were mostly in the typical range of language ability. Consequently, it is unclear whether these findings would look different in children at lower levels of language and/or reading ability who may follow a different developmental trajectory. This paradigm has been used in 16–18 y.o.s with language impairment (LI; McMurray et al., 2014), and LI did not moderate the effect of rStep. This effect is partially replicated here (the effect of composite language on competitor looks), and is distinct from the effects of development we observed. Thus, it is unlikely that younger impaired children would show a different pattern. However, extending this to younger children and to children with dyslexia (which may be linked to phonological deficits, Bishop & Snowling, 2004) remains an important area for research.

The sharpening of phoneme identification

The slope of the phoneme identification function steepens with development (Figure 3). Effects were small but statistically robust in mouse clicks, showing changes between ages 7 and 12 for voicing and fricatives, and continued development through age 18 for voicing. This is consistent with prior studies (Hazan & Barrett, 2000; Nittrouer, 2002; Slawinski & Fitzgerald, 1998), though we further extend the developmental window showing changes in voicing categorization (and possibly fricatives) between 12 and 18. When we examined the dynamic unfolding of categorization (Figure 5, 6, http://osf.io/w5bqg), we saw much bigger effects. Between 7 and 12, there were differences in the ultimate steepness and amplitude of the function (at the end of processing), and the speed at which children achieved this; between 12 and 18, it appeared mostly in the rapidity with which children reached their asymptotic levels.

Thus, even “low level” abilities like speech categorization develop slowly through adolescence. One could dismiss the differences in dynamic identification as mere performance (i.e. children are getting more efficient at accessing knowledge they acquired earlier). However, when we look at the other findings (changes in within category sensitivity, reduction of competitor activation), this under-describes the developmental phenomena. Importantly, speech categorization did not reach asymptotic levels until 1200–1400 msec. Under typical speech rates, by 1200 msec, children might hear 4–5 words—maybe 25 phonemes! As a result, asymptotic behavior is likely a poor descriptor of the speech abilities that children bring to real-world speech perception; rather, performance after only a few

hundred milliseconds might map more closely to real-world performance (c.f., Spivey, 2007). This underscores the importance of examining the automaticity of language processing with dynamic online measures.

The Development of Gradient Categories

Our analysis of within-category was intended to partially isolate the category structure (e.g., the gradient mapping from continuous representations of cues like VOT to phoneme categories) from other factors. It minimized the contribution of trial-by-trial noise in encoding continuous cues (e.g., VOT, frication spectra) by eliminating these "misheard" trials prior to examining the eye-movements[3]. It also isolated development in the overall degree of lexical competition (the main effect of age in Figure 9) from category structure (rStep × age interactions), as lexical competition should affect all steps along the continuum equally. Finally, as the VWP offers a more probabilistic measure (people can look at multiple objects over the trial) that is largely implicit, it was likely to be less sensitive to response-level demands (e.g., to always chose the more active category) that could develop. In this way, these results (e.g., Figure 9) offered a cleaner assessment the internal structure of the categories, the graded mapping between regions of cue-space and categories.

Given this logic, we found robust evidence for differences in within-category sensitivity over development. This appeared to take the form of *less* gradient categories early in development which was most robustly observed for voicing, but also appeared in fricatives. Thus, it appears that as a whole, children develop gradient, within-category activation of lexical competitors over the course of late childhood and early adolescence.

While we attribute this within-category activation to the underlying category structure, there is one alternative hypothesis that was not addressed. At the cue level, it is possible that the issue is not trial-by-trial noise (which our method rules out), but rather the way individual cue values are represented. Consider a model in which cue-values like VOT are represented as a sort of topographic map, with low VOTs at one extreme and high on the other. One source of developmental change might be the precision of encoding on this map (e.g., differences in the "tuning curves" for cues like VOT). For example, in older children, VOTs could be coded highly precisely, with a VOT of 18 msec only activating a narrow region near 18 msec. In contrast, in younger children, they could be more coarsely coded with a VOT of 18 msec partially activating regions corresponding to 12 and/or 24 msec (this would look something like supplementary Figure S2C). This difference would not necessarily alter the final response (since the 18 msec region was the most active in both cases), but could lead to less sensitivity to fine-grained differences since in younger children, the pattern of activity across the VOT map when an 18 msec VOT is heard will partially overlap with the pattern for a 12 msec VOT. This overlap would be seen even if the learned mappings between regions of the cue space and the categories was the same over development. While we cannot rule this out, it is consistent with our broader hypothesis that a crucial development is the functional amount of gradiency in the speech categories, and conversely the functional

[3]We acknowledge it cannot rule out all forms of noise. Listeners could, for example, mishear a 10 msec VOT as 20, but then later noise could still lead to a /b/ response. Moreover, it does not account for noise deep within the category (e.g., at 40 msec of VOT) that does not affect the final response.

sensitivity of the system to fine grained detail. It leaves open the question as to whether this primarily develops via increased precision of cue-encoding (sharpening of the tuning curves), or via changes to category structure.

Regardless of the mechanisms, the *increase* in gradiency is striking given that the interaction took the opposite form in the identification measures. At face value, a shallower identification slope (in younger children) would seem to predict *more* sensitivity to within-category detail. Thus, one might have predicted that children start gradient, but make more confident (sharper) decisions about the category with development. This is consistent with studies like Clayards et al. (2008) that tie both the slope of the categorization function and the degree of gradiency to confidence in the categorization decision. However, this is not what we observed. Over development, children are tuning the structure of the categories to be *more* sensitive to fine-grained differences (more gradient), not more categorical.

Our results are also consistent with a variety of studies in adults that suggest sensitivity to small within-category differences could be beneficial for coping with uncertainty (Clayards et al., 2008; McMurray, Tanenhaus, et al., 2009), for integrating multiple cues (Kapnoula et al., 2017), and for harnessing fine-grained detail for anticipating future events (Gow, 2001). Our data suggests children are learning to deploy these strategies, and at late ages. Ultimately, these strategies allow them to make a more confident (i.e. more categorical) phoneme judgement. Distinct phoneme categories are not just a product of accurately learned boundaries or templates. Rather, distinct categories derive from active perceptual processes by which listener compensate for variability, “explain” the variability in the input, and most importantly here, manage ambiguity (McMurray & Jongman, 2011). These processes go significantly beyond simply matching the input to learned boundaries or templates. The present study suggests that these real-time processes develop late as children learn to solve the problem of acoustic variability.

This suggests that the cause of a shallower identification slope may not be the same as the cause of more gradient within-category responding. Supporting this, Kapnoula et al. (2017) compared the slope from a goodness rating task (tracking gradient responding) with that from a traditional 2AFC phoneme identification task, and found almost no correlation. An analysis of the pattern of variance across the tasks suggested that the 2AFC task may better reflect noise in the signal whereas the rating task reflects the mappings. Thus, the slope of the identification task may ultimately have little to do with the structure of the category.

Lexical Competition

We also saw differences in the amount of lexical competition with age (Figure 9, and Supplement S4 for an analysis of the timecourse). It revealed significant reductions in overall competition between both age contrasts, which cannot be attributed to merely looking around more, as these analyses accounted for differences in unrelated looks. The reduction in competition was accompanied by increases in the speed of activating the target (Supplement S4), similar to prior work (Rigler et al., 2015; Sekerina & Brooks, 2007). It suggests that competition is not simply a product of ambiguity in the signal. Rather, over development, children learn to manage this competition, with a protracted developmental timecourse.

Developmental Processes

Taken as a whole, these results suggest that multiple aspects of speech perception and spoken word recognition are developing throughout childhood and adolescence. What mechanisms might support these changes?

A prominent theory of speech development in infancy is distributional learning, the idea that children use the statistical distribution of cue values to acquire phonological categories (de Boer & Kuhl, 2003; Maye et al., 2003; McMurray, Aslin, et al., 2009). However, this is likely only part of the story. Typical estimates suggest children hear 17,000 words/day (Hart & Risley, 1995); if these are coupled to CHILDES estimates of word frequencies, on a typical day, children likely hear 5250 word-initial stop consonants and 1994 sibilants. By age 7, that's 13.4 million stops and 5.1 million fricatives. Given the robust statistical distributions of these cues, it is unlikely that further input after age 7 is necessary for learning. If statistical learning is insufficient, these changes likely derive from a broader developmental system in which multiple factors in the child and environment contribute.

One possibility may be improvements in ***inhibitory control*** or executive function. Hypothetically, better top-down inhibition could help children suppress competitors and make more discrete decisions about words and phonemes, and adolescence is a time of large changes in executive function and cognitive control (Welsh & Pennington, 1988). While this cannot be ruled out, it may not be directly related to speech. Most models of spoken word recognition do not use domain general inhibition or control processes to suppress competitors; rather inhibition is viewed as a local property within the system (McClelland & Elman, 1986). Supporting this, relationships between speech categorization and executive function are small and often not significant in adults (Kapnoula et al., 2017; Kong & Edwards, 2016). One possibility is that earlier in development, competition is suppressed via cognitive control and over development these become automatized within lexical processing. Alternatively, executive function could play a role in the response system which maps phoneme or lexical activation to the task.

A second factor is the ***lexical growth***. Children learn thousands of words during school age years. This dramatic growth of the lexicon could alter speech perception. A larger lexica could force changes in lexical competition to help make lexical access more efficient. At a fine grained level, the need to distinguish so many words could put pressure on the system to more precisely specify phonological categories (Metsala & Walley, 1998; Walley, Metsala, & Garlock, 2003). Alternatively, the presence of known words in the lexicon could provide a feedback signal to help listeners cope with ambiguous tokens (Feldman et al., 2013); this "clean up" signal could help children refine their categories.

Reading instruction may also be a factor, particularly, during the earlier developmental period studied (7–12). In many reading and spelling curricula, children receive some training in phonological tasks like rhyme judgements, as well as explicit training in letter/sound mappings. This may force children into a more phonemic (rather than holistic) mode of representing the input (Dich & Cohn, 2013). In this way, it may be useful to compare children taught primarily with whole language to those taught with phoneme awareness and phonics.

While lexical growth and reading instruction may account for improvements in speech perception and word recognition as a whole, can they also account for the improvements in children's sensitivity to fine-grained continuous detail (Figure 9)? One possibility is that lexical or orthographic representations serve as an anchor. Lexical representations could help listeners access an idealized auditory representation of the word or phoneme, much in the way that people confronted with an unfamiliar word often ask for it to be spelled. Once the listener knows what a word is supposed to sound like (e.g., it is supposed to have a VOT of 0 msec), this may then permit a more detailed analysis of the signal by computing the difference from the current signal and these idealized expectations (McMurray & Jongman, 2011). It remains to be seen whether this can account for developmental change, but it may represent an avenue by which fine-grained auditory perception can improve with changes in lexical or orthographic knowledge.

Broader Implications

Substantial work on developmental and individual differences is needed to disentangle these hypotheses. However, several concrete findings of the present study constrain such explanations. 1) Speech perception development is a long term process that occurs throughout childhood (and for parallels in speech production, see Sadagopan & Smith, 2008); 2) Children grow more sensitive to fine-grained, gradient detail over development; and 3) Multiple aspects of speech perception are developing in concert with lexical processes. Speech development may derive from a complex developmental system, and not from simple perceptual learning.

The sharpening of identification functions typically observed in development, along with the finding that impaired listeners show shallower identification slopes (Godfrey, Syrdal-Lasky, Millay, & Knox, 1981; Thibodeau & Sussman, 1979; Werker & Tees, 1987) (though see Coady, Evans, Mainela-Arnold, & Kluender, 2007; McMurray et al., 2014) has suggested that gradiency is sub-optimal – the system is "trying" to suppress seemingly irrelevant within-category detail. In contrast, this study (aligning with more recent adult work) shows that the ability to represent and use such gradient information actively develops quite slowly, underscoring the potential functional value of these representations.

This work also has important applied implications. It is well known that meta phonological skills (e.g. phoneme awareness, rhyme awareness) in preschool and kindergarten predict reading outcomes (Parrila, Kirby, & McQuarrie, 2004; Torgesen, Wagner, Rashotte, Burgess, & Hecht, 1997) and training paradigms (Bus & van Ijzendoorn, 1999) and curricula emphasizing phoneme awareness (Ehri et al., 2001) lead to reading gains. Phoneme awareness training, however, is often predicated on the assumption that speech perception and lexical processing is developed by this age. An understanding of the ongoing development of speech, the precise components that are developing, and the mechanisms of development may help shape how phoneme awareness is assessed and taught.

Similarly, work on language impairment (LI) has often stressed the possibility of an auditory or phonological deficit (Bishop & Snowling, 2004; McArthur & Bishop, 2004). However, McMurray et al. (2014) used a nearly identical paradigm with adolescents with and without LI and found only a lexical deficit – there was no rStep × VOT interaction. They suggest a

lexical deficit may be a better characterization of LI than a phonological one. This study extends that by suggesting that such deficits in LI are not merely "delayed development." Younger typical children show a very different profile (McMurray et al., 2010; Rigler et al., 2015). Thus, while the deficits may be primarily lexical, they may also be unique to LI.

Conclusions

The development of speech perception has long been framed in terms of the problem of acoustic variability: how can a child identify the correct boundaries given the variability in the input? However, a close consideration of phonetic data suggests that such boundaries may not be sufficient. Even if a child could find the right boundaries, phonemes overlap heavily, and many tokens will be mis-identified. This problem may be even worse in infant directed speech, which is more variable than adult directed (Cristia & Seidl, 2013; Martin et al., 2015; McMurray, Kovack-Lesh, Goodwin, & McEchron, 2013). We suggest a need to reframe the developmental problem. Children do not need to just identify the categories; they must also develop real-time processing skills to deal with variability, skills like compensating for talker identity and coarticulation (McMurray & Jongman, 2011), managing uncertainty (Clayards et al., 2008), and accessing lexical representations more automatically. As the present study demonstrates, when viewed through the lens of these real-time processes, the development of speech perception may be more protracted and multi-faceted than previously considered.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank Ashley Farris-Trimble for consultation on the design of the study and for generating stimuli; and Tyler Ellis and Claire Goodwin for assistance with data collection. This project was supported by DC0008089 awarded to BM and DC000242 awarded to BM and Bruce Gantz.

References

Andruski JE, Blumstein SE, Burton MW. The effect of subphonetic differences on lexical access. Cognition. 1994; 52:163–187. [PubMed: 7956004]

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B. 1985; 85:289–300.

Bernstein LE. Perceptual development for labeling words varying in voice onset time and fundamental-frequency. Journal of Phonetics. 1983; 11(4):383–393.

Best CT, McRoberts GW, Sithole NM. Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. Journal of Experimental Psychology: Human Perception and Performance. 1988; 14(3):345–360. [PubMed: 2971765]

Bishop DVM, Snowling MJ. Developmental dyslexia and specific language impairment: same or different? Psychological Bulletin. 2004; 130:858–886. [PubMed: 15535741]

Bus AG, van Ijzendoorn MH. Phonological awareness and early reading: A meta-analysis of experimental training studies. Journal of Educational Psychology. 1999; 91(3):403.

Clayards M, Tanenhaus MK, Aslin RN, Jacobs RA. Perception of speech reflects optimal use of probabilistic speech cues. Cognition. 2008; 108(3):804–809. [PubMed: 18582855]

Coady J, Evans JL, Mainela-Arnold E, Kluender K. Children with specific language impairments perceive speech most categorically when tokens are natural and meaningful. Journal of Speech Language and Hearing Research. 2007; 50:41–57.

Cristia A, Seidl A. The hyperarticulation hypothesis of infant-directed speech. Journal of Child Language, FirstView. 2013; :1–22. DOI: 10.1017/S0305000912000669

DahanD, , MagnusonJS. Spoken-word recognition. In: TraxlerMJ, , GernsbacherMA, editorsHandbook of PsycholinguisticsAmsterdam: Academic Press; 2006249283

Dahan D, Magnuson JS, Tanenhaus MK, Hogan E. Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. Language and Cognitive Processes. 2001; 16(5/6):507–534.

de Boer B, Kuhl PK. Investigating the role of infant-directed speech with a computer model. Auditory Research Letters On-Line (ARLO). 2003; 4:129–134.

Dich N, Cohn AC. A review of spelling acquisition: Spelling development as a source of evidence for the psychological reality of the phoneme. Lingua. 2013; 133:213–229.

Dietrich C, Swingley D, Werker JF. Native language governs interpretation of salient speech sound differences at 18 months. Proceedings of the National Academy of Sciences. 2007; 104(41): 16027–16031. DOI: 10.1073/pnas.0705270104

Dollaghan C. Spoken word recognition in children with and without specific language impairment. Applied Psycholinguistics. 1998; 19:193–207.

DunnDM, , DunnLM. Peabody picture vocabulary test: ManualPearson Assessments; 2007

Ehri LC, Nunes S, Willows D, Schuster B, Yaghoub-Zadeh Z, Shanahan T. Phonemic awareness instruction helps children learn to read: Evidence from the National Reading Panel's meta-analysis. Reading Research Quarterly. 2001; 36:250–287.

Eilers RE, Minifie F. Fricative discrimination in early infancy. Journal of Speech and Hearing Research. 1975; 18(1):158–167. [PubMed: 1168827]

Farris-Trimble A, McMurray B. Test-retest reliability of eye tracking in the visual world paradigm for the study of real-time spoken word recognition. Journal of Speech Language and Hearing Research. 2013; 56:1328–1345.

Feldman NH, Griffiths TL, Goldwater S, Morgan JL. A role for the developing lexicon in phonetic category acquisition. Psychological Review. 2013; 120(4):751–778. DOI: 10.1037/a0034245 [PubMed: 24219848]

Fernald A, Perfors A, Marchman VA. Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the 2nd year. Developmental Psychology. 2006; 42(1):98–116. [PubMed: 16420121]

Fry DB, Abramson AS, Eimas PD, Liberman AM. The identification and discrimination of synthetic vowels. Language and Speech. 1962; 5:171–189.

GalleME. PhDUniversity of Iowa; 2014Integration of asynchronous cues in fricative perception in real-time and developmental-time: Evidence for sublexical memory/integration systems.

Galle ME, Klein-Packard J, Schreiber K, McMurray B. What are you waiting for? Real-time integration of cues for fricatives suggests encapsulated auditory memory. submitted.

Galle ME, McMurray B. The development of voicing categories: A meta-analysis of 40 years of infant research. Psychonomic Bulletin and Review. 2014; 21(4):884–906. [PubMed: 24550074]

Godfrey JJ, Syrdal-Lasky AK, Millay KK, Knox CM. Performance of dyslexic children on speech perception tests. Journal of Experimental Child Psychology. 1981; 32:401–424. [PubMed: 7320677]

Gow DW. Assimilation and Anticipation in continuous spoken word recognition. Journal of Memory and Language. 2001; 45:133–139.

HartB, , RisleyT. Meaningful differences in the everyday experience of young American childrenBaltimore, MD: Paul Brookes Publishing; 1995

Hay JF, Graf Estes K, Wang T, Saffran JR. From Flexibility to Constraint: The Contrastive Use of Lexical Tone in Early Word Learning. Child Development. 2015; 86(1):10–22. DOI: 10.1111/cdev.12269 [PubMed: 25041105]

Hazan V, Barrett S. The development of phonemic categorization in children aged 6–12. Journal of Phonetics. 2000; 28(4):377–396. DOI: 10.1006/jpho.2000.0121

Healy AF, Repp BH. Context independence and phonetic mediation in categorical perception. Journal of Experimental Psychology: Human Perception and Performance. 1982; 8(1):68–80. [PubMed: 6460086]

Ishida M, Samuel AG, Arai T. Some people are “More Lexical” than others. Cognition. 2016; 151:68–75. DOI: 10.1016/j.cognition.2016.03.008 [PubMed: 26986746]

Jusczyk PW. From general to language-specific capacities: The WRAPSA model of how speech perception develops. Journal of Phonetics. 1993

Kapnoula E, Winn MB, Kong E, Edwards J, McMurray B. Evaluating the sources and functions of gradiency in phoneme categorization: An individual differences approach. Journal of Experimental Psychology: Human Perception and Performance. 2017; 43(9):1594–1611. [PubMed: 28406683]

Kong EJ, Edwards J. Individual Differences In Categorical Perception Of Speech: Cue Weighting And Executive Function. Journal of Phonetics. 2016; 59

Kuhl PK. Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. Journal of the Acoustical Society of America. 1979; 66:1668–1679. [PubMed: 521551]

Kuhl PK, Conboy BT, Padden D, Nelson T, Pruitt J. Early Speech Perception and Later Language Development: Implications for the “Critical Period”. Language Learning and Development. 2005; 1(3–4):237–264. DOI: 10.1080/15475441.2005.9671948

Kuhl PK, Stevens EHA, Deguchi T, Kiritani S, Iverson P. Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. Developmental Science. 2006; 9:F13–F21. [PubMed: 16472309]

Law F, Mahr T, Schneeberg A, Edwards J. Vocabulary size and auditory word recognition in preschool children. Applied Psycholinguist. 2017; 38(1):89–125.

LuceRD. Individual Choice Behavior: A Theoretical AnalysisNew York: Wiley; 1959

Mainela-Arnold E, Evans JL, Coady J. Lexical representations in children with SLI: Evidence from a frequency manipulated gating task. Journal of Speech Language and Hearing Research. 2008; 51:381–393.

Marslen-Wilson WD, Moss HE, Van Halen S. Perceptual distance and competition in lexical access. Journal of Experimental Psychology: Human Perception and Performance. 1996; 22(6):1376–1392. [PubMed: 8953227]

Martin A, Schatz T, Versteegh M, Miyazawa K, Mazuka R, Dupoux E, Cristia A. Mothers speak less clearly to infants than to adults: A comprehensive test of the hyperarticulation hypothesis. Psychological Science. 2015; 26(3):341–347. [PubMed: 25630443]

Maye J, Werker JF, Gerken L. Infant sensitivity to distributional information can affect phonetic discrimination. Cognition. 2003; 82:101–111.

McArthur GM, Bishop DVM. Which People with Specific Language Impairment Have Auditory Processing Defecits? Cognitive Neuropsychology. 2004; 21(1):79–94. [PubMed: 21038192]

McClelland JL, Elman JL. The TRACE model of speech perception. Cognitive Psychology. 1986; 18(1):1–86. [PubMed: 3753912]

McMurrayB. Nonlinear curvefitting for Psycholinguistics (Version 12.0)2017Retrieved from https://osf.io/4atgv/

McMurray B, Aslin RN. Infants are sensitive to within-category variation in speech perception. Cognition. 2005; 95(2):B15–B26. [PubMed: 15694642]

McMurray B, Aslin RN, Tanenhaus MK, Spivey MJ, Subik D. Gradient sensitivity to within-category variation in words and syllables. Journal of Experimental Psychology, Human Perception and Performance. 2008; 34(6):1609–1631. [PubMed: 19045996]

McMurray B, Aslin RN, Toscano JC. Statistical learning of phonetic categories: Insights from a computational approach. Developmental Science. 2009; 12(3):369–379. [PubMed: 19371359]

McMurrayB, , Farris-TrimbleA. Emergent information-level coupling between perception and production. In: CohnA, FougeronC, , HuffmanM, editorsThe Oxford Handbook of Laboratory PhonologyOxford, UK: The Oxford University Press; 2012

McMurray B, Farris-Trimble A, Seedorff M, Rigler H. The effect of residual acoustic hearing and adaptation to uncertainty in Cochlear Implant users. Ear and Hearing. 2016; 37(1):37–51.

McMurray B, Jongman A. What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. Psychological Review. 2011; 118(2):219–246. [PubMed: 21417542]

McMurray B, Kovack-Lesh K, Goodwin D, McEchron WD. Infant directed speech and the development of speech perception: Enhancing development or an unintended consequence? Cognition. 2013; 129:362–378. [PubMed: 23973465]

McMurray B, Munson C, Tomblin JB. Individual differences in language ability are related to variation in word recognition, not speech perception: Evidence from eye-movements. Journal of Speech Language and Hearing Research. 2014; 57:1344–1362.

McMurray B, Samelson VS, Lee SH, Tomblin JB. Individual differences in online spoken word recognition: Implications for SLI. Cognitive Psychology. 2010; 60(1):1–39. [PubMed: 19836014]

McMurray B, Tanenhaus MK, Aslin RN. Gradient effects of within-category phonetic variation on lexical access. Cognition. 2002; 86(2):B33–B42. [PubMed: 12435537]

McMurray B, Tanenhaus MK, Aslin RN. Within-category VOT affects recovery from “lexical” garden paths: Evidence against phoneme-level inhibition. Journal of Memory and Language. 2009; 60(1): 65–91. [PubMed: 20046217]

MetsalaJL, , WalleyAC. Spoken vocabulary growth and the segmental restructuring of lexical representations: Precursors to phonemic awareness and early reading ability. In: MetsalaJL, , EhriL, editorsWord recognition in beginning literacyMahwah, NJ: Lawrence Erlbaum Associates; 199889120

Miller JL. Internal structure of phonetic categories. Language and Cognitive Processes. 1997; 12:865–869.

Miller JL, Eimas PD. Internal structure of voicing categories in early infancy. Perception & Psychophysics. 1996; 58(8):1157–1167. [PubMed: 8961827]

Miller JL, Volaitis LE. Effect of speaking rate on the perceptual structure of a phonetic category. Perception & Psychophysics. 1989; 46(6):505–512. [PubMed: 2587179]

Narayan CR, Werker JF, Beddor PS. The interaction between acoustic salience and language experience in developmental speech perception: evidence from nasal place discrimination. Developmental Science. 2010; 13(3):407–420. DOI: 10.1111/j.1467-7687.2009.00898.x [PubMed: 20443962]

NeareyTM, , HoganJ. Phonological contrast in experimental phonetics: Relating distributions of measurements in production data to perceptual categorization curves. In: OhalaJ, , JaegerJ, editorsExperimental phonologyNew York, NY: Academic Press; 1986141161

Nittrouer S. Age-related differences in perceptual effects of formant transitions within syllables and across syllable boundaries. Journal of Phonetics. 1992

Nittrouer S. Learning to perceive speech: How fricative perception changes, and how it stays the same. Journal of the Acoustical Society of America. 2002; 112:711–719. [PubMed: 12186050]

Nittrouer S. The role of temporal and dynamic signal components in the perception of syllable-final stop voicing by children and adults. The Journal of the Acoustical Society of America. 2004; 115(4):1777–1790. DOI: 10.1121/1.1651192 [PubMed: 15101656]

Nittrouer S, Miller ME. Predicting developmental shifts in perceptual weighting schemes. Journal of the Acoustical Society of America. 1997; 101:2253–2266. [PubMed: 9104027]

Nittrouer S, Studdert-Kennedy M. The role of coarticulatory effects in the perception of fricatives by children and adults. Journal of Speech, Language, and Hearing Research. 1987; 30(3):319–329.

Parrila R, Kirby JR, McQuarrie L. Articulation rate, naming speed, verbal short-term memory, and phonological awareness: Longitudinal predictors of early reading development? Scientific Studies of Reading. 2004; 8(1):3–26.

Rigler H, Farris-Trimble A, Greiner L, Walker J, Tomblin JB, McMurray B. The slow developmental timecourse of real-time spoken word recognition. Developmental Psychology. 2015; 51(12):1690–1703. [PubMed: 26479544]

Rost GC, McMurray B. Finding the signal by adding noise: The role of non-contrastive phonetic variability in early word learning. Infancy. 2010; 15(6):608.

Sadagopan N, Smith A. Developmental changes in the effects of utterance length and complexity on speech movement variability. Journal of Speech, Language, and Hearing Research. 2008; 51(5): 1138–1151.

Sekerina IA, Brooks PJ. Eye movements during spoken word recognition in russian children. Journal of Experimental Child Psychology. 2007; 98:20–45. [PubMed: 17560596]

SemelEM, , WiigEH, , SecordW. CELF 4: Clinical Evaluation of Language FundamentalsPearson Assessments; 2006

Slawinski EB, Fitzgerald LK. Perceptual development of the categorization of the / r-w / contrast in normal children. Journal of Phonetics. 1998; 26:27–43.

SpiveyMJ. The continuity of mindNew York: Oxford University Press; 2007

Tanenhaus MK, Spivey-Knowlton MJ, Eberhard KM, Sedivy JC. Integration of visual and linguistic information in spoken language comprehension. Science. 1995; 268:1632–1634. [PubMed: 7777863]

Thibodeau LM, Sussman HM. Performance on a test of categorical perception of speech in normal and communication disordered children. Journal of Phonetics. 1979; 7:375–391.

Thiessen ED, Pavlik PI. Modeling the role of distributional information in children's use of phonemic contrasts. Journal of Memory and Language. 2016; 88:117–132. DOI: 10.1016/j.jml.2016.01.003

Torgesen JK, Wagner RK, Rashotte CA, Burgess S, Hecht S. Contributions of phonological awareness and rapid automatic naming ability to the growth of word-reading skills in second-to fifth-grade children. Scientific Studies of Reading. 1997; 1(2):161–185.

Tsuji S, Cristia A. Perceptual attunement in vowels: A meta-analysis. Developmental psychobiology. 2014; 56:179–191. [PubMed: 24273029]

Utman JA, Blumstein SE, Burton MW. Effects of subphonetic and syllable structure variation on word recognition. Perception & Psychophysics. 2000; 62(6):1297–1311. [PubMed: 11019625]

Walley AC, Metsala JL, Garlock VM. Spoken vocabulary growth: Its role in the development of phoneme awareness and early reading ability. Reading and Writing. 2003; 16(1):5–20.

Weber A, Scharenborg O. Models of spoken-word recognition. Wiley Interdisciplinary Reviews: Cognitive Science. 2012; 3(3):387–401. DOI: 10.1002/wcs.1178 [PubMed: 26301470]

WechslerD, , Hsiao-pinC. WASI-II: Wechsler abbreviated scale of intelligencePearson Assessments; 2011

Welsh M, Pennington B. Assessing frontal lobe functioning in children: Views from developmental psychology. Development Neuropsychology. 1988; 4(3):199–230.

Werker JF, Curtin S. PRIMIR: A Developmental Framework of Infant Speech Processing. Language Learning and Development. 2005; 1(2):197–234. DOI: 10.1080/15475441.2005.9684216

Werker JF, Polka L. Developmental changes in speech perception: new challenges and new directions. Journal of Phonetics. 1993; 21:83–101.

Werker JF, Tees RC. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. Infant Behavior and Development. 1984; 7:49–63.

Werker JF, Tees RC. Speech perception in severely disabled and average reading children. Canadian Journal of Psychology. 1987; 41(1):48–61. [PubMed: 3502888]

Zangl R, Klarman L, Thal D, Fernald A, Bates E. Dynamics of Word Comprehension in Infancy: Developments in Timing, Accuracy, and Resistance to Acoustic Degradation. Journal of Cognition and Development. 2005; 6(2):179–208. [PubMed: 22072948]

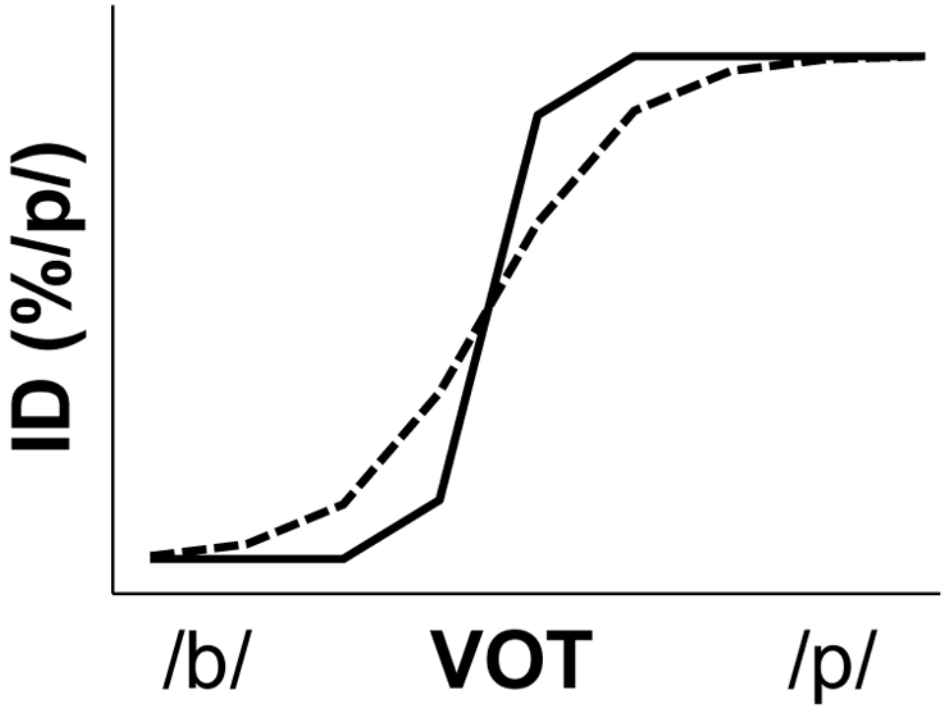

Figure 1.

Schematic results from phoneme decision experiments. Here participants heard tokens from a VOT continua spanning /b/ and /p/ and decided whether each one was /b/ or /p/. Typical studies observe that the slope of the function (at the transition region) gets steeper (more step-like) with age.

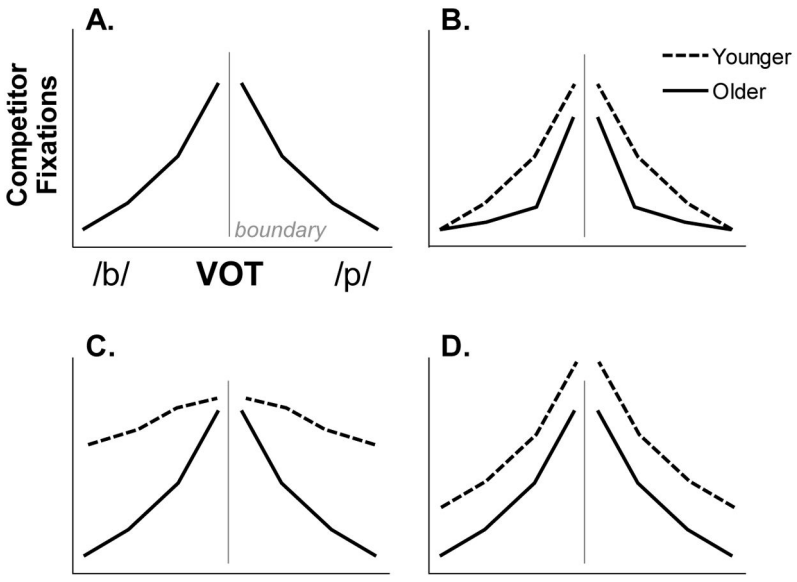

Figure 2.

Schematic results from the McMurray et al. (2002) paradigm. A) Average fixations to the competitor (e.g., /p/ when the stimulus was a /b/) as a function of distance from the participants own boundary; B) Predictions if children become less sensitive to the gradient structure of speech categories (more categorical) with development C) Predictions if children become more sensitive to fine-grained structure with development; D) Predictions if there is no developmental change in how VOT is mapped to categories, but an overall decrease in lexical competition.

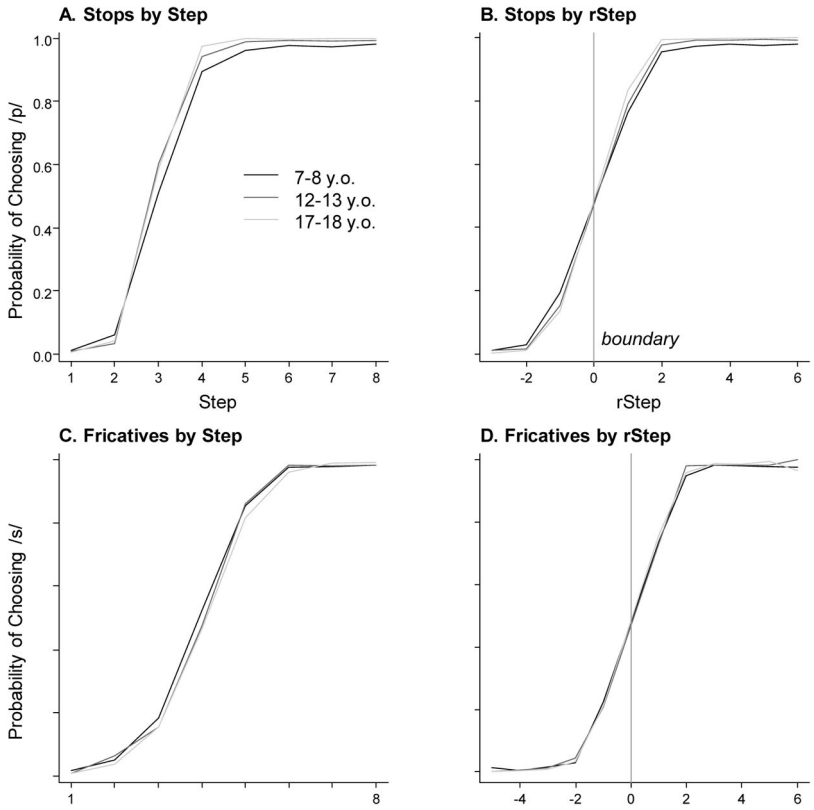

Figure 3.

Identification (mouse click) results. A) Proportion of /p/ responses as a function of step and age for the b/p stimuli; B) Identification of b/p continua as a function of relative step (rStep) which reflects distance from each participants' own category boundary; C) Proportion of /s/ responses as a function of step and age for ʃ/s stimuli; D) ʃ/s identification as a function of rStep.

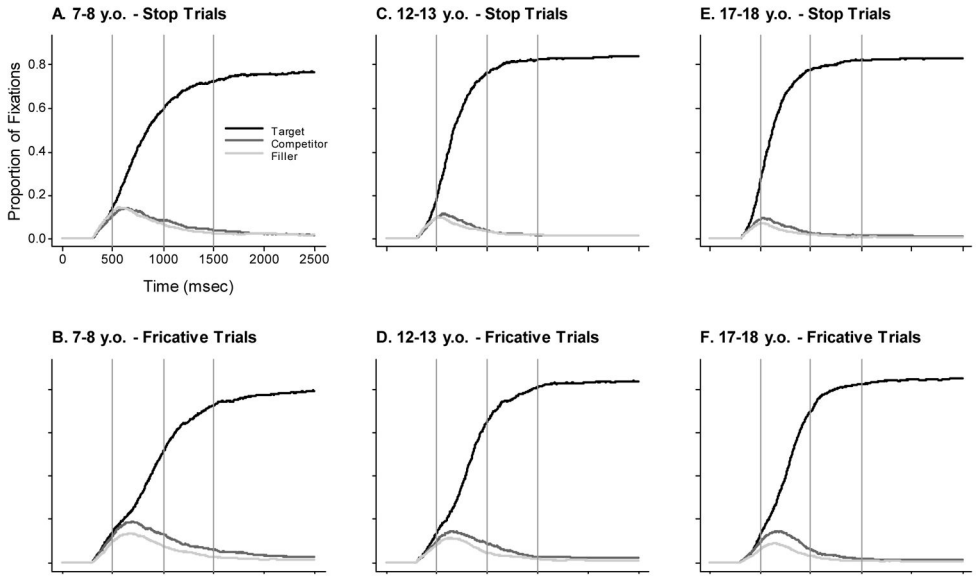

Figure 4.

Proportion fixations to target (the item consistent with the response), the competitor (the other endpoint of the continuum, and the unrelated item as a function of time and age. These figures reflect only the endpoints of the continua.

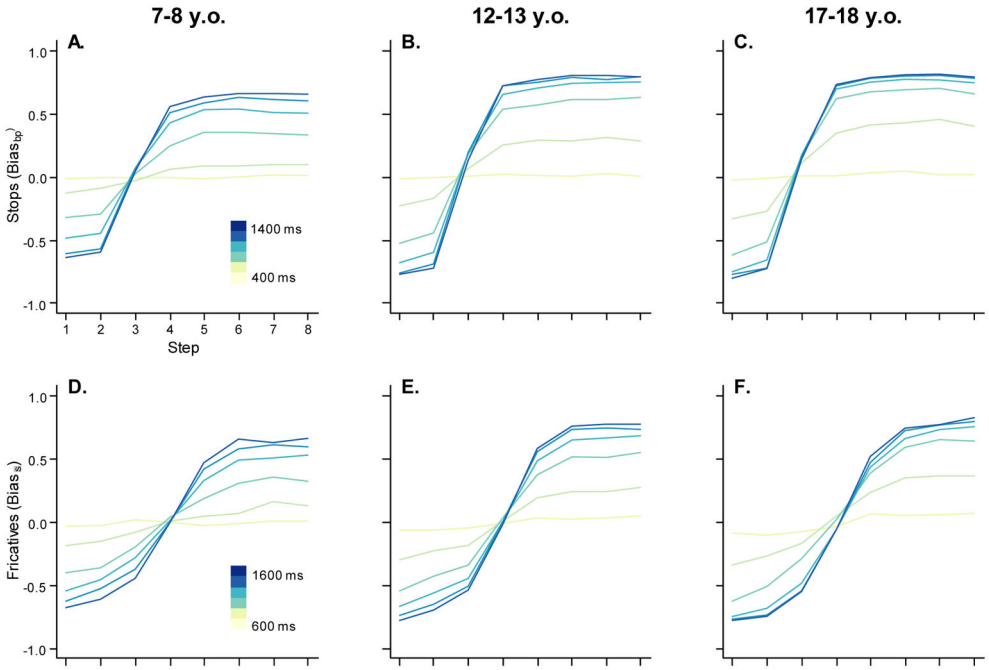

Figure 5.

Categorization functions unfolding over situation time. At each time bin, the bias to fixate /b/ or /p/ (top row) or /ʃ/ or /s/ (bottom row) was computed as a function of continuum step to compute something analogous to a standard identification curve. Each panel shows one continuum at one age group; time is represented by different lines. Animations showing the same data unfolding over time are available at http://osf.io/w5bqg

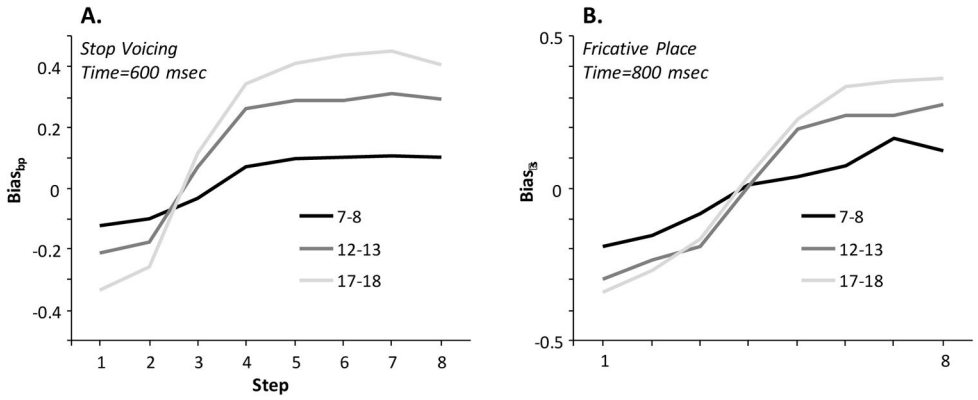

Figure 6.

Bias as a function of continuum step at representative timepoints in processing. A) Stop voicing continuum; B) Fricative place continuum.

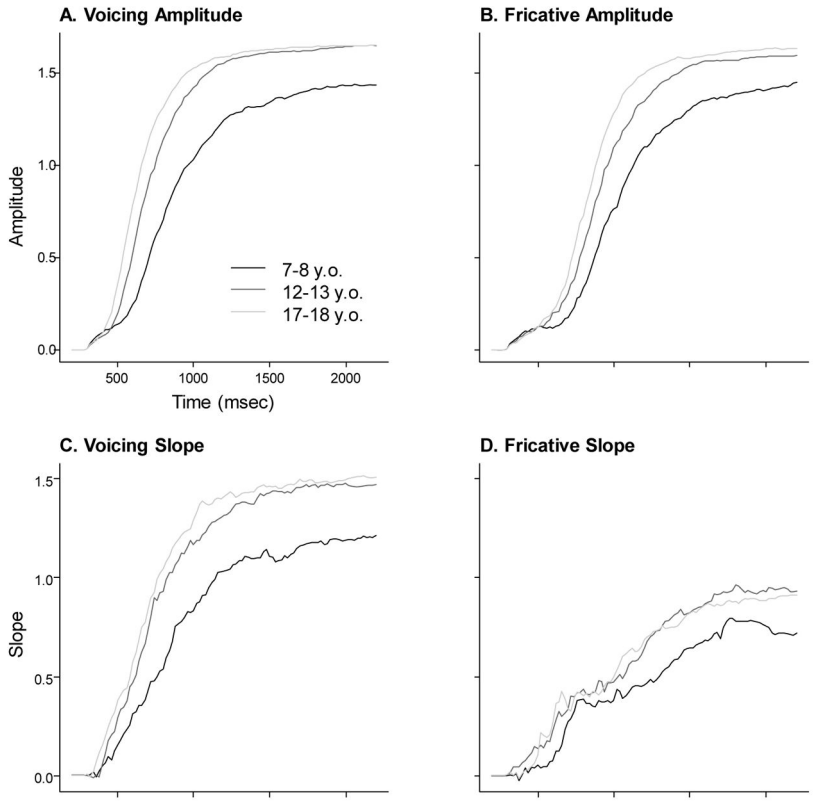

Figure 7.

Properties of the identification curve as a function of time and age. A) Categorization Amplitude (separation between the asymptotes) of voicing identification as a function of time for each age; B) Categorization Amplitude of fricative identification. C) Categorization Slope of voicing identification as a function of time. D) Categorization Slope for fricatives.

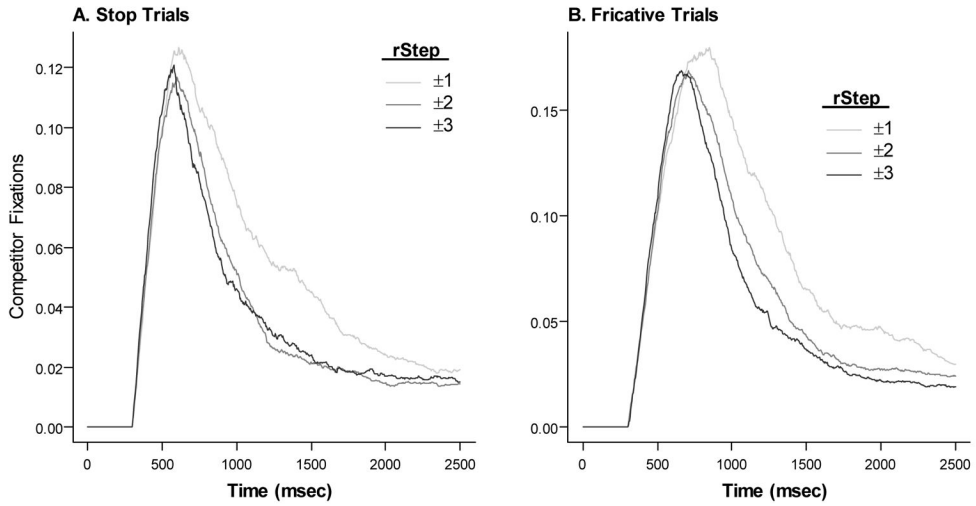

Figure 8.

Looks to the competitor for only trials in which the target was chosen as a function of time and rStep. rSteps of -1 (a /b/ or /ʃ/) were averaged with rSteps of +1 (a /p/ or /s/) and so on. A) For stop voicing continua; B) For fricative place continua

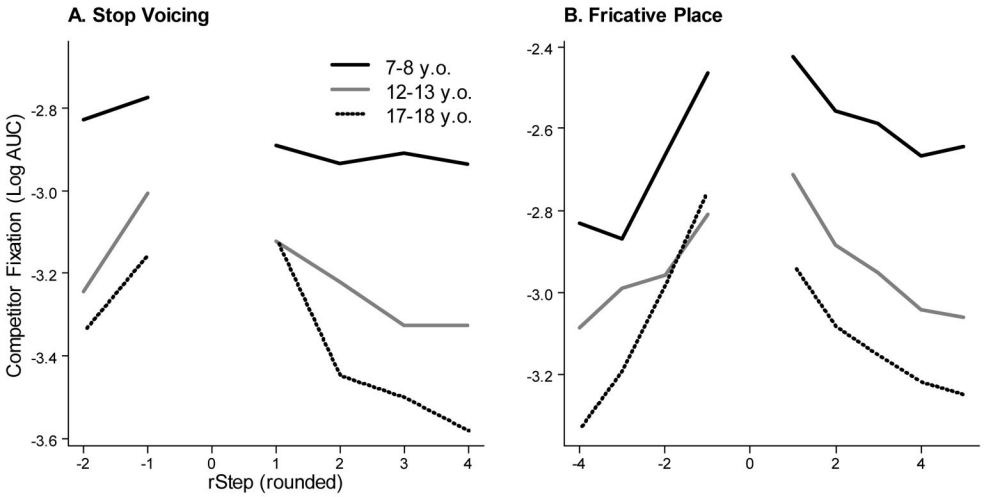

Figure 9.

Area under the Curve (AUC, log scaled) for competitor fixations as a function of rStep and age. rStep was treated as a continuous variable for analysis but rounded here for ease of visualization.

Table 1

Mean Standardized Assessment Scores (standard scores) in each age group. Standard deviations is in parenthesis.

Age-Group (years)	N	PPVT	CELF	WASI
7–8	25	116.9 (11.2)	106.8 (9.5)	108.0 (11.6)
12–13	24	113.3 (15.1)	102.8 (10.1)	95.8 (17.7)
17–18	25	104.8 (11.4)	101.2 (8.7)	99.9 (17.2)

Table 2

Results of a mixed effects model examining identification as a function of Step and age. Separate models were run for stop voicing (top) and fricative place (bottom).

	Effect	B	SE	Z	p
Stop Voicing	Step	2.967	0.138	21.49	<.0001 *
	Young vs. Mid	1.947	0.268	7.28	<.0001 *
	Mid vs. Old	−1.674	0.286	−5.86	<.0001 *
	Language	4.494	1.520	2.96	0.00311 *
	Step × [Young vs. Mid]	1.057	0.130	8.12	<.0001 *
	Step × [Mid vs. Old]	0.949	0.144	−6.62	<.0001 *
	Step × Language	2.442	0.949	2.57	0.01 *
Fricatives Place	Step	2.250	0.149	15.13	<.0001 *
	Young vs. Mid	−0.026	0.154	−0.17	
	Mid vs. Old	0.099	0.151	0.65	
	Language	0.674	1.130	0.60	
	Step × [Young vs. Mid]	0.251	0.102	2.46	.014 *
	Step × [Mid vs. Old]	0.182	0.101	−1.81	.071 +
	Step × Language	2.056	0.769	2.67	.0075 *

* p<.05.

p>.2 not shown.

Table 3

Results of t-tests (p-values) comparing Categorization Amplitude and Slope between adjacent ages. P-values are controlled for False Discovery Rate. –: p>.1. Italics: marginal significance. T-tests have 47 d.f. except where indicated in parenthesis; missing data is due to poor fits for one or more subjects at this time.

	Stop Voicing				**Fricative Place**			
	7–8 vs. 12–13		**12–13 vs. 17–18**		**7–8 vs. 12–13**		**12–13 vs. 17–18**	
Time	**Amp**	**Slope**	**Amp**	**Slope**	**Amp**	**Slope**	**Amp**	**Slope**
500	0.0101 (43)	0.010 (43)	0.0017 (46)	0.097 (46)	- (38)	- (38)	- (38)	- (38)
600	<.0001 (46)	0.0003 (46)	0.0015	-	0.057 (36)	*0.052 (36)*	- (38)	- (38)
700	<.0001	<.0001	0.0088	-	0.0041 (43)	*0.096 (43)*	*0.058 (46)*	- (46)
800	<.0001	<.0001	*0.052*	-	0.0006 (45)	- (45)	*0.052*	-
900	0.0001	0.0013	-	-	0.0007	-	*0.062*	-
1000	0.0003	0.0074	-	-	0.0007	-	*0.078*	-
1200	0.0029	0.038	-	-	0.013	0.023	-	-
1400	0.012	0.032	-	-	0.030	0.022	-	-

Table 4

Summary of t-tests comparing age groups on the parameters that describe how amplitude and steepness change as a function of time.

			7–8 vs. 12–13			12–13 vs. 17–18		
	DV	**Parameter**	**t(47)**	**p**		**t(47)**	**p**	
b/p	Categorization Amplitude	Crossover	6.95	<.001	*	4.54	<.001	*
		Slope	2.46	.018	*	<1		
		Asymptote	2.37	.022	*	<1		
	Categorization Slope	Crossover	<1			<1		
		Slope	1.09	.28		1.63	.11	
		Asymptote	2.27	.028	*	<1		
ʃ/s	Categorization Amplitude	Crossover	3.30	.002	*	2.28	.027	*
		Slope	2.25	.029	*	2.05	.046	*
		Asymptote	1.80	.077	+	<1		
	Categorization Slope	Crossover	<1			<1		
		Slope	1.59	.12		<1		
		Asymptote	2.29	.026	*	<1		

Table 5

Results of a mixed effects model examining competitor fixations as a function of rStep and age for the b/p continuum.

	Effect	B	SE	T	df	p
Voiced (/b/)	rStep	0.167	0.036	4.68	80.2	<.0001 *
	rStep^2	0.124	0.051	2.42	89	.0175 *
	Young vs. Mid	−0.248	0.059	−4.19	81.7	<.0001 *
	Mid vs. Old	−0.174	0.058	−2.99	83.6	.00362 *
	Language	−0.050	0.040	−1.26	71.4	
	rStep × Young vs. Mid	0.106	0.054	1.97	71.9	.052 +
	× Mid vs. Old	0.049	0.054	0.91	75.5	
	× Language	0.017	0.038	0.44	75	
	rStep^2 × Young vs. Mid	0.082	0.073	1.12	78.1	
	× Mid vs. Old	0.053	0.078	0.68	93.8	
	× Language	0.061	0.053	1.14	85.4	
	Unrelated looks	4.863	1.156	4.21	720.5	<.0001 *
Voiceless (/p/)	rStep	−0.042	0.010	−4.16	78	<.0001 *
	rStep^2	0.025	0.007	3.75	152.5	<.0001 *
	Young vs. Mid	−0.372	0.047	−7.99	86.7	<.0001 *
	Mid vs. Old	−0.307	0.045	−6.85	82.1	<.0001 *
	Language	−0.087	0.032	−2.71	75.7	.0084 *
	rStep × Young vs. Mid	−0.033	0.016	−2.07	85.1	.042 *
	× Mid vs. Old	−0.031	0.015	−2.03	78	.046 *
	× Language	0.005	0.011	0.42	78.9	
	rStep^2 × Young vs. Mid	0.020	0.011	1.85	172.8	.067 +
	× Mid vs. Old	0.020	0.010	1.96	155.1	.052 +
	× Language	0.003	0.007	0.45	151.3	
	Unrelated looks	1.130	0.671	1.68	1812.7	.092 +

For p-values,

* p<.05,

† p<.1;

p>.2 not shown.