HHS Public Access

Author manuscript

Clin Linguist Phon. Author manuscript; available in PMC 2017 January 01.

Published in final edited form as:

Clin Linguist Phon. 2017 ; 31(1): 56–79. doi:10.1080/02699206.2016.1233292.

Bias in the Perception of Phonetic Detail in Children's Speech: A Comparison of Categorical and Continuous Rating Scales

Benjamin Munson,

Department of Speech-Language-Hearing Sciences, University of Minnesota, Minneapolis

Sarah K. Schellinger, and

Department of Speech-Language-Hearing Sciences, University of Minnesota, Minneapolis

Jan Edwards

Department of Communication Sciences and Disorders, University of Wisconsin, Madison

Abstract

Previous research has shown that continuous rating scales can be used to assess phonetic detail in children's productions, and could potentially be used to detect covert contrasts. Two experiments examined whether continuous rating scales have the additional benefit of being less susceptible to task-related biasing than categorical phonetic transcriptions. In both experiments, judgments of children's productions of /s/ and /θ/ were interleaved with two types of rating tasks designed to induce bias: continuous judgments of a parameter whose variation is itself relatively more continuous (gender typicality of their speech) in one biasing condition, and categorical judgments of a parameter that is relatively less-continuous (the vowel they produced) in the other biasing condition. One experiment elicited continuous judgments of /s/ and /θ/ productions, while the other elicited categorical judgments. The results of Experiment 1 showed that the influence of acoustic characteristics on continuous judgments of /s/ and /θ/ was stable across biasing conditions. In contrast, the results of Experiment 2 showed that the influence of acoustic characteristics on categorical judgments of /s/ and /θ/ differed systematically across biasing conditions. These results suggest that continuous judgments are psychometrically superior to categorical judgments, as they are more resistant to task-related bias.

The Auditory-Perceptual Assessment of (not-so-)Covert Contrast

The papers in this volume demonstrate persuasively that children are capable of producing systematic phonetic variation that is not easily captured in categorical phonetic transcriptions. The seminal research studies on this topic used acoustic analysis to uncover differences in productions that had been transcribed with the same phonetic symbol, such as Macken and Barton's (1980) study of voice onset time (VOT). Subsequent work has examined this topic with a variety of other instrumental techniques, including direct measures of articulation like electropalatography and ultrasound (e.g., Gibbon, 1999). The inescapable conclusion from these studies is that phonetic acquisition is far more complex than transcription-based studies would suggest. An equally inescapable conclusion is that assessing covert contrast instrumentally is time-consuming, as the measures needed to document it must be taken off-line. Moreover, they rely on our knowledge base of

articulatory-acoustic relationships in the growing vocal tract. That field is very much in its infancy.

An alternative method of assessing covert contrast is to use continuous rating scales to denote category goodness of children's productions. Consider the data in Figure 1. This figure plots the voice-onset time for a hypothetical child's productions of /p/- and /b/-initial words. The data are based on one of the children from Macken and Barton (the subject they nicknamed *Tessa*, taken from session 7). They were estimated from a figure in that paper and match the range, mean, and standard deviation for that participant. Macken and Barton describe Tessa's data in this session as having a *covert contrast*: there is a difference in VOT for /b/ and /p/ targets, but they fall in the range that generally represents adults' perceptual boundaries between /p/ and /b/. Hence, they would not necessarily be transcribed with different symbols. This is shown in the plots of the /p/ and /b/ VOTs in the bottom portion of this figure. Imagine that these productions were presented to listeners in an experiment along with a continuous rating scale in the form of a double-headed arrow anchored by the text "the 'b' sound" and "the 'p' sound." We predict that the distributions of ratings for /b/ and /p/ targets would be different. That is, we predict that these ratings would track the variation in VOT more finely than the categorical transcriptions would. This is shown in the hypothetical distributions of continuous ratings above the continuous rating scale at the top of Figure 1. A continuous rating scale, then, could differentiate among cases of covert contrast (like that shown in this figure) and the case of a true substitution error. Binary phonetic transcriptions cannot distinguish between these cases.

Our speculation about how listeners might perceive the productions analyzed by Macken and Bartonis based on the results of numerous studies that elicit continuous ratings of children's productions of sounds (Julien & Munson, 2012; McAllister Byun, Halpin, & Harel, 2015; Munson & Brinkman, 2004; Munson, Johnson, & Edwards, 2012; Munson, Schellinger, Edwards, Beckman, & Meyer, 2010; Strömbergsson, Savli, & House, 2015). These studies have shown that listeners are capable of providing continuous ratings of children's productions, provided that the productions themselves vary continuously in how closely they resemble canonical productions of sounds. This is consistent with studies of adult speech perception showing that listeners are capable of perceiving phonetic detail in speech sounds. These include behavioral studies utilizing explicit ratings, behavioral studies using eye-tracking, and neurophysiological investigations. Taken together, they show that listeners can access within-category phonetic detail during speech perception, and that this detail affects both the encoding of acoustic signals and their parsing into phonetic categories (Carney, Widin, & Viemeister, 1977; Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Massaro & Cohen, 1983; McMurray, Tanenhaus, Aslin, & Spivey, 2003; Miller, 1994; Toscano, McMurray, Dennhardt, & Luck, 2010).

This use of rating scales may be particularly important for describing the speech of children with communication disorders. For example, Todd, Edwards, and Litovsky (2011) found that /s/ productions of children with cochlear implants that were transcribed as correct by trained phoneticians were systematically acoustically different from /s/ productions of children with normal hearing of the same chronological age. Revai (2016) observed similar findings for /k/ productions by children with cochlear implants and their normal hearing age

peers. While acoustic measures such as those used in the Todd et al. and Revai studies are difficult to obtain and interpret, continuous ratings scales are relative quick and easy to administer. Indeed, Bernstein, Todd, and Edwards (2013) used a continuous rating scale to evaluate adult judgments of the stimuli from the Todd et al. (2011) paper and found that adults rated /s/ productions by children with cochlear implants as less /s/-like than productions by a normal hearing comparison group. This was true despite the fact that only productions transcribed to be correct were examined. Similar findings are reported by Strombergsson et al. (2015), who showed that the /t/ productions of children with speech sound disorders are perceived as less /t/-like than those of their peers with typical speech development, even though these children's /t/ productions were transcribed as accurate. Thus, a growing body of research suggests that continuous rating scales provide results that are reliable and may yield additional information about children's productions that is similar to what can be gleaned from acoustic analysis.

The current investigation follows up on our earlier studies of ratings of children's speech by examining their susceptibility to bias. Speech perception is subject to myriad biases. The same sound can be labeled differently depending on a variety of social, pragmatic, and linguistic factors. As argued by Kent (1996), these biases almost certainly affect the way that speech is perceived in clinical assessments. Indeed, the fact that the speech of individuals with communication impairments is atypical suggests that its perception might be subject to even stronger biases than the perception of typical speech, as listeners generally have less experience perceiving disordered speech, and hence have weaker expectations of how it should sound. Kent provides no prescription to remediate these biases, other than to develop conscious awareness of them in hopes that awareness might mitigate their influence on assessments. Another possibility is that different *types* of ratings are more or less susceptible to bias. In this paper, we explore the possibility that continuous ratings are less easily biased than are categorical phonetic transcriptions. Our hypothesis is based on the widely acknowledged principle that speech perception involves sensory perception (i.e., the uptake of continuous acoustic and visual information) and categorization (i.e., parsing phonetic events as members of a small number of linguistic units, such as phonemes). The categorization process necessarily abstracts away from the variation that is present in the signal. The phonetic characteristics of different tokens of a unit (i.e., tokens of the phoneme /s/) vary considerably based on numerous factors, such as phonetic context and speaker identity. We reason that drawing people's attention to the different sources of variation in the signal will affect categorizations. For example, in a task devised to draw attention to speaker-specific variation in fricatives, a decision to categorize a given token as /s/ or /θ/ might be more strongly affected by a listener's knowledge of how these sounds vary from talker to talker. In a task devised to draw attention to vowel context, a decision to categorize the very same token as /s/ or /θ/ might be more strongly affected by a listener's knowledge of how these sounds vary across vowel contexts. The continuous rating task does not require the same degree of abstraction as the categorization task; hence, we predict that continuous ratings (for example, of the degree to which a given sound resembles /s/ or /θ/) will not differ across conditions that draw attention to different sources of variation.

This paper reports on a set of experiments designed to test this hypothesis. These experiments interleave continuous judgments of the extent to which tokens resemble

prototypical /s/ or /θ/ (Experiment 1) or categorical judgments of whether a given token is an instance of /s/ or /θ/ (Experiment 2) with two other types of judgments. In one condition, judgments of /s/ and /θ/ are interleaved with categorical judgments of the vowel the child produced. In the other condition, judgments of /s/ and /θ/ are interleaved with continuous judgments of the gender typicality of the child's voice. We predict that the categorical judgments in Experiment 2 will be more easily biased (i.e., they will differ across conditions) than will the continuous ratings in Experiment 1.

Attention and Speech Perception

Given the topic of this paper, we provide a brief review of published literature on the effect of attention on speech perception. This topic has been studied numerous ways. Some studies have examined the effect of divided attention on speech perception, while others have examined individuals' ability to selectively attend to a sound or a phonetic feature. One simple way attention has been studied has been to examine whether listeners can selectively attend to a position in a word when monitoring whether a sound occurred. Pitt and Samuel (1990) used this tactic and found that listeners were readily able to focus on a particular part of a word when monitoring for a phoneme, even when the phoneme-monitoring task was made especially difficult by paring it with a grammatical classification task.

Another way that attention in speech perception has been studied is by exploiting the fact that phonemic contrasts are often implemented via multiple acoustic and visual parameters. For example, the English voicing contrast in word-initial plosives is conveyed primarily through voice-onset time, but also through the f0 of the following vowel at onset and the intensity of the burst. Listeners' attention to different cues to the voicing contrast has been found to shift in conditions of divided attention (Gordon, Eberhardt, & Rueckl, 1993). Listeners can actively direct their attention to different cues. For example, Francis, Kaganovich, and Driscoll-Huber (2008) found that listeners can be trained to attend to different cues to word-initial consonant voicing in English. Chandrasekaran, Yi, Smayda, and Maddox (2015) found that English-speaking listeners' attention could be drawn to different acoustic cues when learning Mandarin lexical tones. Idemaru and Holt (2014) showed that listeners' attention to different cues to word-initial voicing in English could be changed by varying the extent to which they correlated in a stimulus set.

Attention can also be modulated by social factors. Janson and Schulman (1983) found that Swedish listeners' perception of the /æ/-/ɛ/ distinction depended on whether or not they were primed to think that they were listening to a language variety in which the contrast was neutralized. That is, attention to phonetic variation could be modulated by introducing expectations about whether sounds should be the same or different. More recently, Munson et al. (2010) demonstrated that listeners' tendency to label children's fricative productions as /s/ or /θ/ in a speech-perception experiment depended on whether or not the word "lisp" was used in the instructions for the experiment. Stimuli that were labeled as /θ/ when the word "lisp" was not used in the instructions were more likely to be labeled as /s/ when the word "lisp" was used. This suggests that the ways that listeners map acoustic variation onto phonetic categories differs depending on their expectations about how the talkers speak. These results are broadly consistent with a host of studies showing that listeners calibrate

their expectations about linguistic variation depending on presumed attributes about a speaker (i.e., Hay & Drager, 2010; Johnson, Strand, & D'Imperio, 1999; Munson, 2011).

Gradiency in Speech Perception

Our study of bias gives us an opportunity to test an additional, ancillary hypothesis about the nature of continuous ratings of children's speech. Specifically, it allows us to examine possible reasons why listeners vary in the extent to which their responses are continuous when presented with a continuous rating scale. Continuous rating scales are only useful insomuch as they elicit a continuous response to continuously varying stimuli. That is, such scales are of limited utility if listeners' responses are clustered at the endpoints of the scale, or at a discrete and small number of locations on the continuous scale. Such response patterns simply replicate categorical systems like phonetic transcription.

Theoretically, listeners could vary completely from those who are *completely continuous* (i.e., listeners whose ratings are equally distributed along the scale) to those who are *completely categorical* (i.e., listeners whose ratings are limited to the two endpoints of the scale). Data presented by Schellinger, Munson, and Edwards (2016, this volume), show that this is rarely the case. While some listeners' ratings clustered at a small number of locations on the continuous scale that Schellinger et al. used, most listeners' ratings were distributed along the scale in meaningful ways. However, listeners in that study did vary considerably in the extent to which their ratings were distributed across the visual analog scale used to elicit them. We will refer to the distribution of ratings along this scale as *gradiency of response*. Listeners' response patterns in Schellinger et al. were such that they could be called *more-gradient* or *less-gradient*. Consider the two listeners in Figure 2. These data come from two listeners in Schellinger et al. The listener whose response distributions are plotted on the top panel of this figure is less gradient: there are local modes in the response distribution at the two endpoints. The listener whose response distributions are plotted on the bottom panel of this figure is more gradient: the responses are roughly evenly distributed across the scale.

Experiment 1 can test, indirectly, one hypothesis for why listeners differ in the overall gradiency of their response. Specifically, we can test the hypothesis that listeners vary in how continuously they rate children's speech production because they are attending to different types of information in the speech signal. Those who respond less gradiently are hypothesized to be attending to linguistic units (such as a phoneme, or an attribute of a speaker) that contrast with one another in a less-gradient manner, while those who respond more gradiently are hypothesized to be attending to linguistic units that contrast with one another in a more-gradient manner. Consider two contrasts among linguistic units in children's speech: the acoustic contrast between two vowels, and the difference in gender typicality between male and female children's voices. Children produce vowels correctly (as judged by adult native speaker transcription) relatively early in life. This suggests that the acoustic characteristics of pairs of vowels are well separated. Indeed, published data confirm this (Chung, Kong, Edwards, Weismer, Fourakis, & Hwang, 2012). In contrast, the difference in perceived gender typicality of male and female children's voices is more poorly separated. Perry, Ohde, and Ashmead (2001) showed that average gender typicality ratings of 4-year-old boys' and girls' voices are statistically significantly different, a finding

that was replicated by Munson and Baylis (2007). However, in Munson and Baylis's data, there was considerable overlap in ratings for boys' and girls' voices. This is consistent with a host of other studies on gender typicality within biological sex, across the lifespan (Munson, 2007; Munson, Crocker, Pierrehumbert, Owen-Anderson, & Zucker, 2015).

There has been little examination of systematic differences among listeners in how continuous or categorical their speech perception is. It is important to point out that the vast majority of speech perception experiments use methods that lend themselves to very different measures of gradiency from those shown in Figure 2. Most speech perception experiments use continua of sounds that vary in a small number of acoustic parameters. Listeners label stimuli along those continua as one of two categories. Logistic regression analysis is used to predict judgments from step number along the continuum. When the stimuli are presented multiple times, listeners who are inconsistent in their judgments have identification functions that are shallower than listeners who are consistent. These shallow identification functions are sometimes described as *less categorical*. Using this method, it has been found that the extent to which phonetic identification is categorical increases with age. The degree of categoricity in children's speech perception increases over the first decade of life, and is not yet fully adult-like at age 10 (Hazan and Barrett, 2000). Moreover, children with language-based communication problems are often observed to be less categorical than their typically developing peers (Manis et al., 1997; Rvachew & Jamieson, 1989). These results suggest that highly categorical speech perception is an index of skill. This interpretation is also supported by some research on adult speech perception. Munson, Johnson, and Edwards (2012) found that highly trained speech-language pathologists perceive children's speech more categorically than do phonetically untrained listeners.

However, there is also evidence in the literature that more-gradient perception of the type shown in Figure 2 is associated with *superior* performance on some tasks. Kong and Edwards (2011) found that native English-speaking listeners whose perception of stop consonant voicing resembles the more-gradient listener in Figure 2 are better able to attend to a secondary cue to voicing (f0), than were listeners whose perception resembles the less-gradient listener. Kapnoula, McMurray, and Edwards (2015) found that more gradient listeners, relative to less gradient listeners, are better able to recover from transient misidentification of stop voicing in an eye-tracking task using the visual world paradigm. Simply put, there is no consensus in the research literature on the causes and consequences of gradiency in perception.

In Experiment 1of this paper, we examined whether we could induce listeners to rate children's speech more or less gradiently depending on whether their attention was drawn to a factor that was inherently more-gradient or inherently less-gradient. We predicted that judgments of /s/ and /θ/ would be more gradient when they were interleaved with judgments of the gender typicality of children's speech than when interleaved by judgments of the vowel the child produced. That is, we predicted that the other judgment with which we interleaved the fricative judgments would induce a more-gradient or less-gradient mode of perception of fricatives. We reinforced the inherent degree of contrast between vowel and gender typicality by always pairing the less-gradient contrast with a categorical judgment (in this case, paring the less-gradient vowel contrast with a task in which listeners chose which

vowel the child produced from a set), and the more-gradient contrast with a continuous rating (in this case, paring the more-gradient gender contrast with a task in which listeners provided a continuous rating of the gender-typicality of children's voices). If we found that listeners' judgments were more gradient when interleaved with judgments of gender, then we could make recommendations concerning the best methods for making gradient judgments of children's speech. We might also speculate that more-gradient listeners in previous studies (like those in Schellinger et al., shown in Figure 2) were those who habitually attend to gradient information in speech.

Experiment 1: Continuous Ratings of Fricatives

Methods

Participants—The participants for Experiment 1 were 19 adults (14 women, 5 men) aged 18 to 45. They were recruited via fliers and word-of-mouth in a large university community. Most were students or staff at the university. They all reported being native speakers of a North American variant of English (defined as acquiring English from birth from at least one parent who was a native speaker of a North American variety of English) with no past or current speech, language, or hearing impairment. They were compensated $10 for their participation.

Stimuli—Stimuli for this experiment were derived from 200 productions of /s/- and /θ/-initial words by 43 different 2- to 5-year-old children (21 girls, 22 boys). The stimuli were collected by having a sample of typically developing children produce words that begin with /s/ or /θ/. These were elicited in a picture-prompted auditory word repetition task, in which children viewed a picture on a computer screen while hearing an audio prompt of its name. Children were asked to repeat each audio prompt. The stimuli included both real words and nonwords, the latter being paired with novel objects. Nonwords were used to elicit a sufficient number of responses for target /θ/, which has a low type frequency in English. The stimuli were elicited in roughly five different vowel contexts: high-front, mid-front, low-central, mid-back, and high-back. The data were collected as part of a cross-linguistic study of consonant acquisition in which these productions were compared to productions of children acquiring a variety of other languages (Edwards & Beckman, 2008). Hence, the vowel contexts were chosen to be ones that would be as equivalent as possible across a variety of languages with different vowel systems, including the five-vowel systems of Greek and Japanese and the seven-vowel system of Korean.

Children's productions were transcribed as [s], [θ], or as a sound intermediate between them (following the suggestion of Stoel-Gammon, 2001). The intermediate sounds were classified as being either closer to [s] (denoted here as [s:θ]) or closer to [θ] ([θ:s]). The stimuli were produced by a variety of boys and girls across the age range. We chose the stimuli to sample correct productions and the different error types as evenly as possible, including as many possible combinations of targets and transcribed categories. This resulted in six different transcription categories: [s] for /s/, [s] for /θ/, [s:θ], [θ:s], [θ] for /s/, and [θ] for /θ/. These were the same stimuli that have been used in a variety of previous studies and which are

described in greater detail elsewhere (i.e., Munson et al., 2010; Schellinger et al., 2016 [this volume]).

Six acoustic measures were taken for each stimulus. These were used both as predictors of listeners' responses in the experiment and to describe the stimuli beyond what is possible with transcription. Three of these were measures of the fricatives, two were measures of the vowels, and one compared the fricative to the vowel. The first two measures were the first (M1, sometimes referred to as *Centroid Frequency*) and second (M2) spectral moments of a 40 ms interval of frication noise centered at the fricative midpoint. The first spectral moment is important for distinguishing /s/ from /ʃ/ in the productions of normal adults, and the second spectral moment is important for distinguishing /s/ from /θ/. The third measure was the duration of the fricative. This measure was included because we observed that the stimuli transcribed as [θ] were shorter than those transcribed as [s]. The next two measures were of the vocalic portion of the stimuli. The first of these was the second formant frequency of the vowel at its onset. As shown by Li, Edwards, and Beckman (2009) and Li et al. (2011), this measure distinguishes between children's productions of /s/ and /ʃ/, and predicts adults' judgments of how /s/- or /ʃ/-like a fricative is. The next measure was of the f0 of the following vowel at midpoint. This measure was included because previous research has shown that the f0 of a vowel in a fricative-vowel sequence influences whether listeners judge the fricative as /s/ or /θ/ (Munson & Coyne, 2010). The final measure was the difference in intensity between the fricative and the following vowel. This measure was included because previous research has shown that /θ/ is less intense than /s/ (e.g., Jongman et al., 2000). The methods for annotating the productions and extracting the acoustic measures were identical to previous studies of the παιδολογος corpus, as described in Li, Beckman, and Edwards (2009) and Li (2012). Readers should note that the onset F2 values were hand measured, and that the f0 was estimated using the pitch track filter in Praat (Boersma, 2001). Outliers were hand-measured.

A summary of the acoustic measures is given in Table 1, which provides statistics separately for the six transcription categories: [s] for/s/, [s] for /θ/, [s:θ], [θ:s], [θ] for /s/, and [θ] for /θ/. The six acoustic measures were used as predictors in a series of stepwise linear discriminant function analyses (DFA). The first DFA predicted whether the sound was transcribed as [θ] (including [θ:s] productions) or [s] (including [s:θ] productions). The second predicted whether the sound was transcribed as [s], [θ], or as either of the intermediate categories. The third DFA predicted membership in one of the six transcription categories. The M1, M2, and relative intensity significantly improved all three models' categorization accuracy. In addition, duration improved the second model's categorization accuracy.

Procedure—Each participant was tested individually in a sound-proof booth, seated in front of a computer monitor. Each of the 200 CV stimuli was played over headphones in random order using E-Prime software (Schneider, Eschmann, & Zuccolotto, 2002). The design of this experiment is shown schematically in Figure 3. There were two blocks of the experiment. In both blocks, listeners were informed that they would hear consonant-vowel syllables taken from words that were supposed to start with "s" or "th." Instructions gave examples of words beginning with /θ/ to cue listeners that they were to listen for /θ/ rather

than /ð/. In one block of the experiment (labeled the *Vowel Block* in Figure 3), listeners were told that on a given trial, they would either rate the goodness of the consonant, or judge the vowel category that the child produced. For the consonant-goodness trials, the listeners were asked to rate the consonant in each CV syllable using a visual analog scale (shown in the top panel of Figure 4) that was presented on the computer monitor. Listeners were explicitly instructed to click the location along the line that corresponded with the percept of proximity to “s” or “th”, and were encouraged to use the entire line. For the vowel-category trials, the listeners were told to identify the vowel that the child produced from a set of five possible vowels (shown in the bottom panel of Figure 3): “ee” (corresponding to /i/), “ey” (/eɪ/), “aa” (/ɑ/), “oh” (/oʊ/), or “oo” (/u/). The instructions included keywords that contained these vowels. The orthographic representations chosen were those most frequently provided by a group of undergraduate students who were asked to give a two-letter spelling of these five vowels. Listeners responded in the vowel-category trials by pressing one of five keys on a computer keyboard.

In the other block of the experiment (referred to as the *Gender Block* in Figure 3), listeners were told that on a given trial, they would either rate the goodness of the consonant, or rate the gender typicality of the child’s voice. The consonant-goodness trials employed the same method as described above. In the gender-typicality trials, listeners were presented with a visual-analog scale (shown in the middle panel of Figure 4) that was presented on the computer monitor. They were explicitly instructed to click the location along the line that corresponded to their judgment of how boy-like or girl-like the child sounded. They were encouraged to use the entire line. In both blocks, listeners did not know which rating they would be providing (consonant goodness, vowel category, gender typicality) until after the stimulus had finished playing. This was done to prevent participants from calibrating their attention to the stimulus based on what they would be rating.

The 200 stimuli were divided into two groups of 100 stimuli. The groups of stimuli were balanced to include the same distribution of talkers, transcribed consonants, and vowel targets. In each block, 100 stimuli were used for consonant-goodness ratings and the other 100 were used for the other rating, either vowel category or gender typicality. Across the entire experiment, consonant-goodness ratings were provided for all 200 stimuli (i.e., the 100 stimuli used for vowel-category judgments in the vowel block were used for consonant-goodness judgments in the gender block). Within each block, the order of stimuli was fully randomized. The order of blocks was randomized across subjects.

Data Analysis—The click location for each stimulus trial was analyzed in terms of the number of pixels along the x-dimension of the visual analog scale. The left end of the VAS line (corresponding to “the ‘s’ sound”) was denoted as the zero point and the right end of the VAS line corresponded with 535 pixels. Any clicks that fell off the line in the horizontal dimension were assigned these minimum and maximum values (i.e., clicks left of the line were assigned “0” and clicks right of the line were assigned 535). All responses that were more than +/− 25 pixels from the line in the y-dimension were excluded from the analysis.

For ease of interpretation, click locations for each trial were transformed to a measure indicating their location in terms of the proportion of the line. The click location for each

trial was divided by the maximum value of 535, resulting in click location values that ranged from zero to one. These were then inverted, so that click locations closer to zero correspond with percepts more like “the ‘th’ sound” and click locations closer to one correspond with percepts of more like “the ‘s’ sound”. The inversion was done so that the ratings in this paper would be comparable to the ratings in previous reports of listeners’ perception of these stimuli, in which the ratings were inverted for reasons not directly relevant to this paper. A click location of 0.5 indicates that the listener perceived the sound as completely intermediate between /s/ and /θ/.

Results

Prior to conducting analyses of the fricative ratings, the gender typicality ratings were analyzed qualitatively. In particular, we assessed whether the gender-typicality ratings utilized the entire visual analog scale. A finding that the ratings were clustered at the endpoints of the scale would indicate that listeners were treating this task like a binary rating task. The distribution of ratings for individual listeners is shown in Figure 6. As this figure shows, the ratings were indeed distributed across the entire scale. No listener had a bimodal distribution of ratings. Hence, we can be confident that the gender rating task indeed elicited a continuous rating.

Predictors of Individual Listeners’ VAS ratings—In the first analysis, a series of linear mixed-effects models was built to examine whether listeners gave different average ratings to the sounds in the different transcription categories, and whether the effect of stimulus acoustics on ratings differed systematically across conditions. These analyses addressed our primary research question concerning whether ratings of /s/ and /θ/ were different in the vowel and gender blocks, i.e., whether they were biased by condition. The dependent measure was the VAS rating. The package *lme4* (Bates, Mächler, Bolker, & Walker, 2014) in the R statistical environment was used. To assess statistical significance of effects within models, the *lmerTest* package was used (Kuznetskova, Brockhoff, & Christensen, 2015). To assess the significance of a factor in a model, we assessed whether the model with the factor had a better fit than a model without that factor. Following the suggestion of Barr, Levy, Scheepers, and Tily (2013), we included all logically possible random slopes for each new fixed effect that we added.

In the first set of analyses, we began by building a base model that contained only an overall intercept and random intercepts for stimuli and listeners. A model that included a factor for transcription category (with all six levels) improved the fit of this model, $\chi^2_{[df=5]} = 255.36$, $p < 0.001$. This model was re-run using all possible reference levels for the factor, so that we could examine whether all pairwise differences among levels were significantly different. These analyses confirmed that the ratings for each of the transcription categories were indeed different from all of the others. Adding a binary term dummy-coding condition (using contrast coding) did not improve model fit, either alone or in an interaction with the transcription category factor, $\chi^2_{[df=6]} = 7.53$, $p = 0.28$. That is, there was no effect of condition (gender block vs. vowel block) on VAS ratings.

The next set of analyses examined the effect of acoustic characteristics on ratings. It began with the same base model. The second model included predictors for all six acoustic measures described in Table 2. These were z-transformed for the analysis. Adding these predictors improved the fit of the model significantly $\chi^2_{[df=6]} = 103.11$, $p < 0.001$. Adding a predictor contrast-coding condition, either alone or in interaction with any of the acoustic measures, did not improve model fit, $\chi^2_{[df=76]} = 6.78$, $p = 0.45$. The result of the less complex model is shown in Table 4. As this table shows, all acoustic measures except onset F2 were significant predictors of listeners' ratings.

Individual Listener VAS Response Patterns—The next set of analyses examined the shape of individual listeners' response distributions. These analyses address our ancillary research question of whether gradiency of response differs as a function of condition, and whether responses are more gradient in the gender block than in the vowel block. Only the consonant-goodness trials were examined. We examined whether these differed as a function of the block in which the ratings were made. Our first analysis examined the extent to which listeners in the consonant-goodness task used the entire line (like the more-gradient listener in Figure 2), or whether their ratings clustered at discreet locations along the line (like the less-gradient listener in Figure 2). Specifically, we examined the extent to which individual listeners' ratings differentiated among the six transcription categories described earlier: [s] for /s/, [s] for /θ/, [s:θ], [θ:s], [θ] for /s/, and [θ] for /θ/. To examine this, we conducted a series of one-way ANOVAs predicting VAS rating from the six transcription categories. Post-hoc Scheffe tests were used to determine the number of homogeneous subsets of ratings that were present in the data. Homogenous subsets are clusters of levels of a categorical independent variable which do not differ from one another, but which do differ from variables in other subsets. For example, one listener might have three homogeneous subsets, one of which comprises the ratings [s] for /s/ and [s] for /θ/, one of which comprises the ratings for [s:θ] and [θ:s], and one of which comprises the ratings for [θ] for /s/ and [θ] for /θ/. We reasoned that more-gradient listeners will have more homogeneous subsets of ratings in their data. This analysis was used previously to study gradiency of response by Schellinger et al. (2016, this volume).

The number of homogeneous subsets for the gender block was either 1 (3/19 listeners), 2 (8/19 listeners), 3 (6/19), or 4 (2/19). The number of homogeneous subsets for the vowel block was either 1 (4/19 listeners), 2 (7/19 listeners), 3 (4/19), or 4 (4/19). The specific subsets that were generated for each listener differed; however, no listener had a homogeneous subset comprising two non-adjacent transcription categories in the order [s] for /s/ > [s] for /θ/ > [s:θ] > [θ:s] > [θ] for /s/ > [θ] for /θ/. Examples from two participants are shown in the middle row of Figure 5. The top row of Figure 5 plots histograms of these listeners' distributions of fricative ratings, and shows that both listeners resemble the more-gradient listener from Figure 2. The listener in the left column has four homogeneous subsets: one for sounds transcribed as [s] regardless of the target, one for sounds transcribed as either of the intermediate categories, one for sounds transcribed as [θ] for target /s/ (i.e., as the commonly occurring 'frontally misarticulated /s/' error), and one for sounds transcribed as [θ] for target /θ/. The listener whose data are in the right column has three homogeneous subsets of data: one for sounds transcribed as [s] regardless of the target, one

for sounds transcribed as either of the intermediate categories, and one for sounds transcribed as [θ] regardless of the target.

A non-parametric Wilcoxon signed ranks test examined whether the number of subsets for the 19 listeners differed as a function of condition. The test was not significant, $z = -0.378$, $p = 0.71$. Hence, there was no evidence that the gradiency of response differed as a function of condition.

The second analysis used mixture models to analyze the shape of individual listeners' response distributions. Mixture models are a class of analyses that decompose complex, multimodal distributions into component, unimodal distributions. Each component distribution's mean and variance is provided, along with the percentage of ratings that are comprised by that component distribution. This analysis is based solely on the shape of the distribution of ratings. It does not consider any of the acoustic or perceptual characteristics of the stimuli. The mixture modeling reported in this paper used the *densitymclust* function from the R package *mclust* (Fraley & Referty, 2002). The algorithm uses an optimization procedure to determine the number of underlying distributions comprising a target distribution. Our analysis specified the shape of the distributions to be Gaussian. Given that the *densitymclust* algorithm is somewhat anticonservative, we wanted to restrict the number of underlying distributions. Hence, we specified the maximum number of distributions to be 8. We reasoned that more-gradient listeners would have more component distributions of ratings. This analysis was used previously to study gradiency of response by Schellinger et al. (2016, this volume).

In the gender block, the number of distributions was either 2 (10/19), 3 (4/19), 4 (1/19), 5 (3/19), or 8 (1/19). The number of distributions in the vowel block was either 1 (1/19), 2 (7/19), 3 (6/19), 4 (2/19), or 5 (3/19). An illustration of the outcome of the mixture modeling procedure is shown in the bottom row of Figure 5. The response distribution for the listener on the left was decomposed into three component distributions, with mean click locations of 0.034, 0.200, and 0.664, variances of 0.0004, 0.0072, and 0.0253, and which comprised 13%, 45% and 42% of the distribution, respectively. The response distribution for the listener on the right was decomposed into five distributions, with mean click locations of 0.003, 0.097, 0.280, 0.663, and 0.953, variances of 0.0001, 0.0016 0.0112, 0.0162, and 0.0002, and which comprised 8%, 23%, 30%, 34%, and 5% of the distribution, respectively.

The number of component response distributions for each subject's responses in the vowel and gender blocks was compared using a Wilcoxon signed ranks test. The test was not significant, $z = -0.229$, $p = 0.82$. Hence, there was again no evidence that the gradiency of response differed as a function of condition.

Discussion

The results of this experiment suggest that continuous ratings of consonant goodness are relatively impervious to biasing, at least of the type introduced in this experiment. The average ratings given to different categories were statistically equivalent in these two conditions, and the weighting given to different acoustic measures was similar across these two conditions. Moreover, the shape of response distributions did not differ systematically

between a condition in which listeners were biased to respond more-gradiently (by attending to a more-continuous variable, gender typicality, and rating it using a continuous rating scale), and one in which they were biased to respond less-gradiently (by attending to a more-categorical variable, vowel category, and rating it using a categorical judgment).

One interpretation of the negative findings of this experiment is that reliable, valid continuous ratings of children's speech can be made on a variety of tasks, including ones in which attention is divided between ratings of accuracy and ratings of a variety of different parameters of children's speech. This would be a very desirable outcome, as assessments of children's speech are made in a variety of settings, and are often made as part of assessments of other characteristics of speech. To examine this possibility further, we conducted a second experiment using the same biasing method, but instead looking at binary judgments of fricative place of articulation. If biasing is found in a binary judgment experiment, then this strengthens our argument for the use of continuous rating scales. Moreover, a finding of biasing in that experiment would allay any concern that the novel, previously untested method we used in Experiment 1 was not effective at inducing bias in speech perception overall.

Experiment 2: Binary Ratings

Methods

Participants—The participants for Experiment 2 were 21 adults aged 18 to 45 (16 women, 5 men). They were recruited via fliers and word-of-mouth in a large university community. Most were students or staff at the university. They all reported being native speakers of a North American variety of English (defined as acquiring English from birth from at least one parent who was a native speaker of a North American variety of English) with no past or current speech, language, or hearing impairment. They were compensated $10 for their participation.

Stimuli—The stimuli were the same as those for Experiment 1.

Procedures—The procedures were the same as those for Experiment 1, with one exception. Rather than clicking on a continuous rating scale to judge the fricatives, listeners clicked on one of two boxes on a screen, marked either "the 's' sound" or "the 'th' sound." The choice to have listeners click on a box rather than using the keyboard was so that the two experiments would be as similar as possible in the devices that they used. That is, if Experiment 2 elicited readings via the keyboard, then any differences between Experiments 1 and 2 might be attributable to the devices being used rather than to different cognitive processes. The experiment was designed so that the listener had to click within the box to proceed to the next trial. This ensured that there would be no lost trials.

Data Analysis—The location that the listener clicked on the screen was converted into a binary judgment, depending on whether they clicked on the box labeled "the 's' sound" or "the 'th' sound."

Results

As in Experiment 1, we examined qualitatively whether the gender-typicality ratings were continuous prior to conducting any further analyses. The distribution of ratings for individual listeners is shown in Figure 7. As this figure shows, the ratings were indeed distributed across the entire scale. No listener had a bimodal distribution of ratings. Hence, we can be confident that the gender rating task indeed elicited a continuous rating, as was the case in Experiment 1.

Our next analyses focused on whether the fricative judgments differed systematically as a function of the condition in which they were made. A series of logit mixed-effects models examined whether condition affected performance. Logit models were used because the outcome in Experiment 2 was binary, i.e., whether the listener judged the sound to be /s/ or /θ/. The same procedures for model fitting were used as in Experiment 1. As with Experiment 1, the first model examined whether condition interacted with transcription category. Here, a model including a six-level factor for transcription category fitted the data significantly better than a baseline model with only random intercept ($\chi^2_{[df=26]} = 3095.2$, $p < 0.001$). However, a model that included a term for condition did not improve model fit beyond that, either alone or in an interaction with the transcription category factor ($\chi^2_{[df=6]} = 6.91$, $p = 0.33$). The best-fitting model was re-run with each level of the transcription factor variable serving as the reference level. This served as a post-hoc test to assess whether all pairwise differences between transcription categories were significantly different. This series of models showed that the binary judgments differed significantly across all six transcription categories. The results of this analysis suggest that the distribution of binary judgments among the six transcription categories did not differ as a function of condition.

The next set of models examined whether condition interacted with acoustic predictors. As in the analyses for Experiment 1, the acoustic measures were z-transformed before being added to the model. The model with acoustic predictors in it fit the data better than the model without ($\chi^2_{[df=12]} = 2398$, $p < 0.001$). The next model included an interaction between condition (coded with contrast coding) and each of the acoustic predictors. This model fit the data significantly better than the model without condition ($\chi^2_{[df=7]} = 20.33$, $p = 0.005$). The outcome of this model is shown in Table 3, which displays only those factors and interactions that were significant in the model. As this table shows, condition interacted significantly with two acoustic predictors, M1 (i.e., centroid frequency) and the f0 of the vowel at midpoint. These interactions are shown in Figures 8 and 9, which plot the functions relating these two predictors to judgments in the two conditions. Figure 8 shows the interaction between M1 and condition. It shows that fricative judgments were more strongly affected by M1 in the gender condition than in the vowel condition. This might be due to listeners strongly attending to M1 as a cue to gender typicality in children's speech in the gender block. This is consistent with previous research showing that listeners use fricative M1 as a cue to the gender typicality of children's voices (Munson et al., 2015).

Figure 9 shows that the interaction between condition and midpoint f0 occurred because there was a qualitatively different influence of f0 on fricative judgments in the two conditions. In the vowel block, higher vowel f0 was associated with more /θ/ judgments. This is consistent with the findings of Munson and Coyne (2010). In the gender block, the

relationship was the opposite. Separate models were run for responses in the two conditions. These confirmed that the influence of midpoint f0 on responses was significant in both models. However, the effect was much stronger in the gender block than in the vowel block. The reason for the asymmetry in the direction of the effect is unclear. However, there is a plausible explanation for the difference in the strength of the relationships. In the gender block, listeners may have attended more to vowel f0 than they did in the vowel block because they were using f0 as a cue to the gender typicality of children's speech. This, too, is consistent with previous research (Munson et al., 2015).

The results thus far suggest that condition biased the binary responses in Experiment 2 (though only in the analysis of acoustic predictors of ratings), but not the continuous judgments in Experiment 1. However, there is the possibility that the difference between Experiments 1 and 2 was not in the biasing *per se*, but in the models used to fit the data. The models used to examine the data in Experiment 1 were linear models, as the fricative ratings were continuous.

The models used to examine data in Experiment 2 were logistic models, as the fricative judgments were binary. To examine the possibility that the differences between the experiments were due purely to statistical techniques, we reanalyzed the data from Experiment 1 using a logistic model. Given that logistic models are only appropriate for binary data, we transformed the continuous ratings from Experiment 1 to binary judgments, based on whether the ratings were made on the /s/ or /θ/ side of the continuous rating scale. These binary judgments were the dependent measures in a series of models in which the acoustic variables were predictors. The same iterative model-building procedure described above was used. The model with acoustic predictors improved on the base model ($\chi^2_{[df=12]} = 1298.6$, $p < 0.001$). However, adding an interaction with condition did not improve this model ($\chi^2_{[df=7]} = 4.01$, $p = 0.78$). The coefficients for the best-fitting model are shown in Table 4. This shows a pattern of significance that mirrors that from Experiment 1: all of the acoustic measures except onset F2 frequency predict judgments.

Discussion

The results of Experiment 2 indicated that judgments of whether a sound is /s/ or /θ/ could be biased by whether these judgments were interleaved with judgments of the vowel that the child produced, or ratings of the gender typicality of the child's voice. Specifically, the conditions induced listeners to weight acoustic characteristics of stimuli differently. This finding stands in contrast to Experiment 1, in which condition did not influence continuous judgments of how /s/- or /θ/-like a sound is. This finding is important for two reasons. The first reason is that it demonstrates that the lack of an effect in Experiment 1 was not due to the experimental manipulation of interleaving judgments of fricatives with judgments of vowels or gender being ineffective at changing individuals' judgments of fricatives. Rather, it suggests that this manipulation can affect the perception of fricatives, but only when those responses are binary judgments, and not when they are continuous ratings. Moreover, the effect is only seen in analyses of the weighting given to different acoustic parameters when making binary judgments. The specific cognitive mechanism that explains the differences in biasing across the two tasks is outside the scope of this investigation. However, from a

purely descriptive standpoint it suggests that the biasing happens not in the encoding of acoustic variation, but in the stage of processing in which acoustic variation is weighted to make a categorization decision. The lack of an effect of condition on weighting in Experiment 1 may be because the task did not require phonemic categorization.

General Discussion

The two experiments in this paper examined the susceptibility of judgments of children's speech to different types of biasing. The specific types of biasing that were examined were ones that we thought would be encountered in real-world assessments of children's speech. In real-world assessments, individuals often make judgments of different aspects of children's speech concurrently or in sequence. For example, clinicians may be transcribing multiple consonants as correct or incorrect at the same time that they are assessing voice quality. Across the two experiments, our results suggest that continuous ratings are less susceptible to bias than categorical ones are. The results of this investigation provide evidence-based recommendations for the use of continuous rating scales in assessing children's speech. An ancillary purpose was to examine whether differences in the gradiency of response differed as a function of the attentional demands of the task. The results of Experiment 1 did not help us understand why individuals vary in the extent to which their continuous ratings of children's speech are fully gradient. While there were measurable differences in the degree to which individual responses were continuous in Experiment 1, these did not vary systematically between conditions. Our search for predictors of individual differences in the gradiency of response is, therefore, ongoing. One promising set of measures has emerged from work by Kong and Edwards (2011) and Kapnoula et al. (2015). Findings in those studies suggest that these individual differences in gradiency of response are accompanied by differences in how much listeners attend to a secondary cue (such as f0 in the case of differentiating the stop voicing contrast), and these individual differences may also be related to differences in cognitive control.

This study had several limitations. One of these is the general paucity of research on effective methods for manipulating attention during speech perception. Without such a literature, it is unclear whether the results of this study will generalize to other conditions of divided or focused attention. A second of these is the lack of a measure of inter-rater reliability. Unlike in previous research (i.e., Schellinger et al., 2016), we did not measure intra-rater reliability. Our motivation was to control the length of the experiment. However, intra-rater reliability may have differed across conditions. These weaknesses aside, the results of the studies in this paper provide further evidence that continuous rating scales are useful auditory-perceptual measures for the assessment of covert contrast.

Acknowledgments

We thank Emma Hage, Eden Kaiser, and Mara Logerquist with help testing participants in this study. We thanks Fangfang Li and Jeffery Holliday for help with stimulus selection and acoustic analysis. We thank Mary E. Beckman for her input on this work as it was in progress. The authors report no conflicts of interest.

References

Barr D, Levy R, Scheepers C, Tily H. Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language. 2013; 68:255–278.

Bates, D.; Maechler, M.; Bolker, B.; Walker, S. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7. 2014. http://CRAN.R-project.org/package=lme4

Bernstein, S.; Todd, A.; Edwards, J. How do adults perceive the speech of children with cochlear implants?. Poster presented at the Annual Conference of the American Speech-Language-Hearing Association; Chicago, IL. 14–16 November; 2013.

Carney AE, Widin G, Viemeister N. Noncategorical perception of stop consonants differing in VOT. Journal of the Acoustical Society of America. 1977; 62:961–970. [PubMed: 908791]

Chandrasekaran B, Yi HG, Smayda KE, Maddox WT. Effect of explicit dimensional instruction on speech category learning. Attention, Perception, & Psychophysics. 2016; 78:566–582.

Chung H, Kong E, Edwards J, Weismer G, Fourakis M, Hwang Y. Cross-linguistic studies of children's and adults' vowel spaces. Journal of the Acoustical Society of America. 2012; 131:442–454. [PubMed: 22280606]

Clayards M, Tanenhaus M, Aslin R, Jacobs R. Perception of speech reflects optimal use of probabilistic cues. Cognition. 2008; 108:804–809. [PubMed: 18582855]

Cutler A, Mehler J, Norris D, Segui J. Phoneme identification and the lexicon. Cognitive Psychology. 1987; 19:141–177.

Edwards J, Beckman ME. Methodological questions in studying consonant acquisition. Clinical Linguistics and Phonetics. 2008; 22:937–956. [PubMed: 19031192]

Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association. 2002; 97:611–631.

Francis A, Kaganovich N, Driscoll-Huber C. Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English. Journal of the Acoustical Society. 2008; 124:1234–1251.

Gibbon F. Undifferentiated lingual gestures in children with articulation/phonological disorders. Journal of Speech, Language, and Hearing Research. 1999; 42:382–397.

Gordon P, Eberhardt J, Rueckl J. Attentional modulation of the phonetic significance of acoustic cues. Cognitive Psychology. 1993; 25:1–42. [PubMed: 8425384]

Hazan V, Barrett S. The development of phonemic categorization in children aged 6–12. Journal of Phonetics. 2000; 28:377–396.

Hay J, Drager K. Short-term exposure to one dialect affects processing of another. Language and Speech. 2010; 53:447–471. [PubMed: 21313989]

Idemaru K, Holt L. Specificity of dimension-based statistical learning in word recognition. Journal of Experimental Psychology: Human Perception and Performance. 2014; 40:1009–1021. [PubMed: 24364708]

Janson T, Schulman R. Non-distinctive features and their use. Journal of Linguistics. 1983; 19:321–336.

Johnson K, Strand E, D'Imperio M. Auditory–visual integration of talker gender in vowel perception. Journal of Phonetics. 1999; 27:359–384.

Jongman A, Wayland R, Wong S. Acoustic characteristics of English fricatives. Journal of the Acoustical Society of America. 2000; 108:1252–1263. [PubMed: 11008825]

Julien H, Munson B. Modifying speech to children based on their perceived phonetic accuracy. Journal of Speech, Language, and Hearing Research. 2012; 55:1836–1849.

Kapnoula, E.; McMurray, B.; Edwards, J. Gradient Categorization of Speech Sounds Helps Listeners Recover from Lexical Garden Paths. Poster presented at the 14th Auditory Perception, Cognition, and Action Meeting; November 19; Chicago, IL. 2015.

Kong, E.; Edwards, J. Individual differences in speech perception: Evidence from visual analogue scaling and eye-tracking. Proceedings of the 17th International Congress of Phonetic Sciences; 2011. p. 1126-1129.

Kuznetsova, A.; Brockhoff, P.; Christensen, R. lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). R package version 2.0-0. 2013. http://CRAN.R-project.org/package=lmerTes

Li F. Language-specific developmental differences in speech production: a cross linguistic acoustic study. Child Development. 2012; 83:1303–1315. [PubMed: 22540834]

Li F, Edwards J, Beckman ME. Contrast and covert contrast: The phonetic development of voiceless sibilant fricatives in English and Japanese toddlers. Journal of Phonetics. 2009; 37:111–124. [PubMed: 19672472]

Li F, Munson B, Edwards J, Yoneyama K, Hall K. Language specificity in the perception of voiceless sibilant fricatives in Japanese and English: Implications for cross-language differences in speech-sound development. Journal of the Acoustical Society of America. 2011; 129:999–1011. [PubMed: 21361456]

Macken M, Barton D. The acquisition of the voicing contrast in English: a study of voice onset time in word-initial stop consonants. Journal of Child Language. 1980; 7:41–74. [PubMed: 7372738]

Manis FR, McBride-Chang C, Seidenberg MS, Keating P, Doi LM, Munson B, Peterson A. Are speech perception deficits associated with developmental dyslexia? Journal of Experimental Child Psychology. 1997; 66:211–235. [PubMed: 9245476]

Massaro D, Cohen M. Categorical or continuous speech perception: a new test. Speech Communication. 1983; 2:15–35.

McAllister Byun, T.; Halpin, P.; Harel, D. The Scottish Consortium for ICPhS. Crowdsourcing for gradient ratings of child speech: Comparing three methods of response aggregation. Proceedings of the 18th International Congress of Phonetic Sciences; Glasgow, UK: University of Glasgow; 2015. Paper retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0935.pdf on March 17, 2016

McMurray B, Tanenhaus M, Aslin R, Spivey M. Probabilistic constraint satisfaction at the lexical/phonetic interface. Journal of Psycholinguistic Research. 2003; 32:77–97. [PubMed: 12647564]

Miller J. On the internal structure of phonetic categories: a progress report. Cognition. 1994; 50:271–285. [PubMed: 8039364]

Munson B. The acoustic correlates of perceived sexual orientation, perceived masculinity, and perceived femininity. Language and Speech. 2007; 50:125–142. [PubMed: 17518106]

Munson B. The influence of actual and imputed talker gender on fricative perception, revisited. Journal of the Acoustical Society of America. 2011; 5:2631–2634.

Munson, B.; Baylis, A. Gender Typicality in the Speech of Children with Phonological Disorder. Poster Presentation at the Symposium for Research on Child; 2007.

Munson B, Brinkman KN. The effect of multiple presentations on judgments of children's speech production accuracy. American Journal of Speech-Language Pathology. 2004; 13:341–354. [PubMed: 15719900]

Munson B, Coyne A. The influence of apparent vocal-tract size, contrast type, and implied sources of variation on the perception of American English voiceless lingual fricatives. Journal of the Phonetic Society of Japan. 2010; 14:48–59.

Munson B, Crocker L, Pierrehumbert J, Owen-Anderson A, Zucker K. Gender typicality in children's speech: A comparison of the speech of boys with and without gender identity disorder. Journal of the Acoustical Society of America. 2015; 137:1995–2003. [PubMed: 25920850]

Munson B, Edwards J, Schellinger SK, Beckman ME, Meyer MK. Deconstructing phonetic transcription: covert contrast, perceptual bias, and an extraterrestrial view of *Vox Humana*. Clinical Linguistics and Phonetics. 2010; 24:245–260. [PubMed: 20345255]

Munson B, Johnson J, Edwards J. The role of experience in the perception of phonetic detail in children's speech: A comparison of speech-language pathologists with clinically untrained listeners. American Journal of Speech-Language Pathology. 2012; 24:124–139.

Pitt M, Samuel A. Attention allocation during speech perception: how fine is the focus? Journal of Memory and Language. 1990; 29:611–632.

Reidy, P.; Kristensen, K.; Winn, M.; Litovsky, R.; Edwards, J. The acoustics of word-initial fricatives and their effect on word-level intelligibility in children with bilateral cochlear implants. 2016. Manuscript submitted for publicationlearningtotalk.org

Revai, D. Undergraduate honor's thesis. UW-Madison; 2016. Production of stop contrasts in children with normal hearing and with cochlear implants. available at learningtotalk.org

Rvachew S, Jamieson D. Perception of voiceless fricatives by children with a functional articulation disorder. Journal of Speech and Hearing Disorders. 1989; 54:193–208. [PubMed: 2709838]

Schneider, W.; Eschmann, A.; Zuccolotto, A. E-Prime user's guide. Pittsburgh, PA: Psychology Software Tools, Inc; 2002.

Strömbergsson S, Salvi G, House D. Acoustic and perceptual evaluation of category goodness of /t/ and /k/ in typical and misarticulated children's speech. Journal of the Acoustical Society of America. 2015; 137:3422–3435. [PubMed: 26093431]

Toscano J, McMurray B, Dennhardt J, Luck S. Continuous perception and graded categorization: Elecrophysiological evidence for a graded relationship between the acoustic signal and perceptual encoding of speech. Psychological Science. 2010; 21:1532–1540. [PubMed: 20935168]

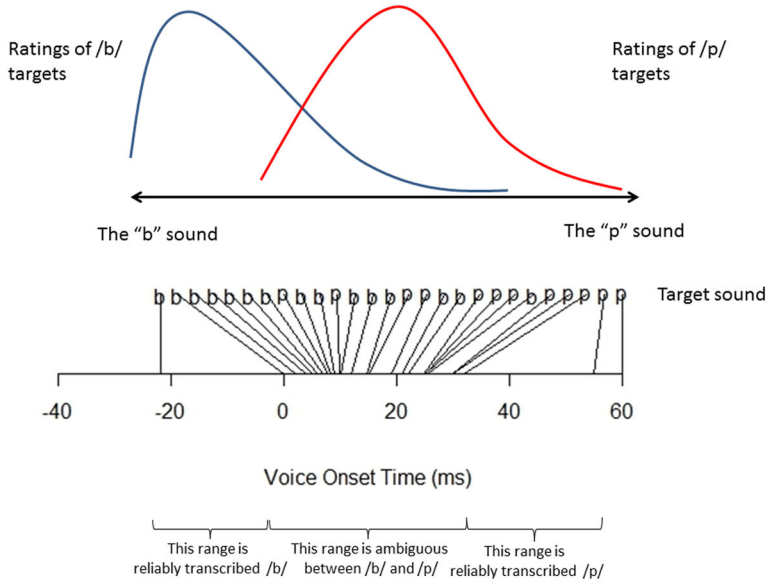

Figure 1.

An illustration of covert contrast in VOT, adapted from Macken and Barton (1980)

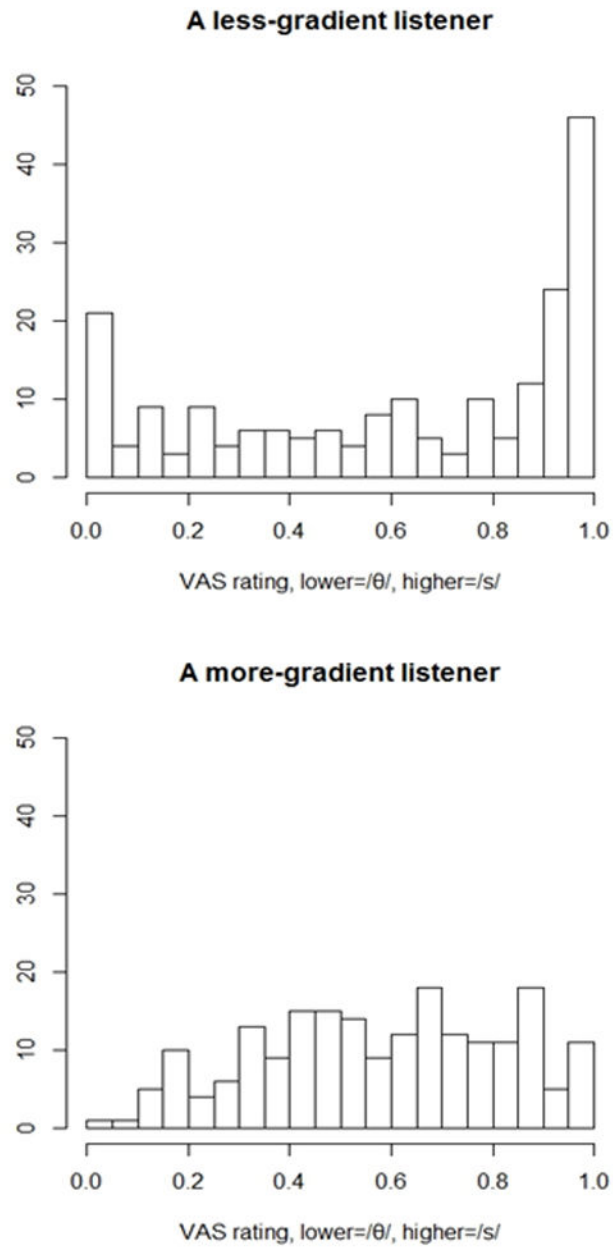

Figure 2.

Two listeners' performance on a continuous-rating scale task, chosen to illustrate a less-gradient listener (top) and a more-gradient listener (bottom)

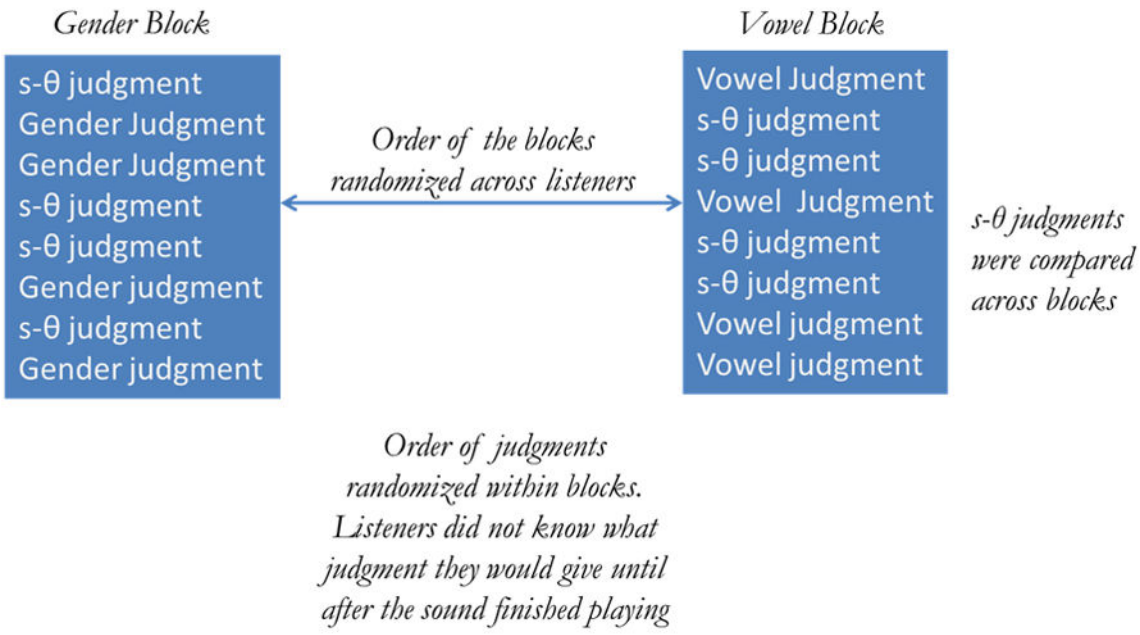

Figure 3.
Schematic of the design of Experiment 1.

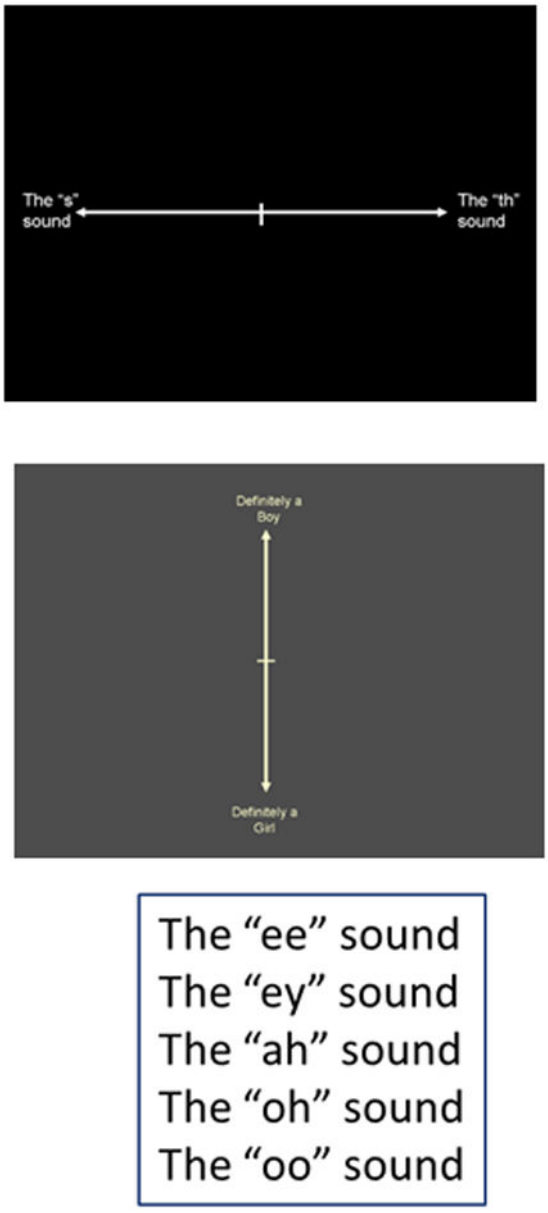

Figure 4.
Response displays used in Experiment 1. Top panel: display used to elicit fricative ratings. Middle panel: display used to elicit gender-typicality ratings. Bottom panel: display used to elicit vowel judgments.

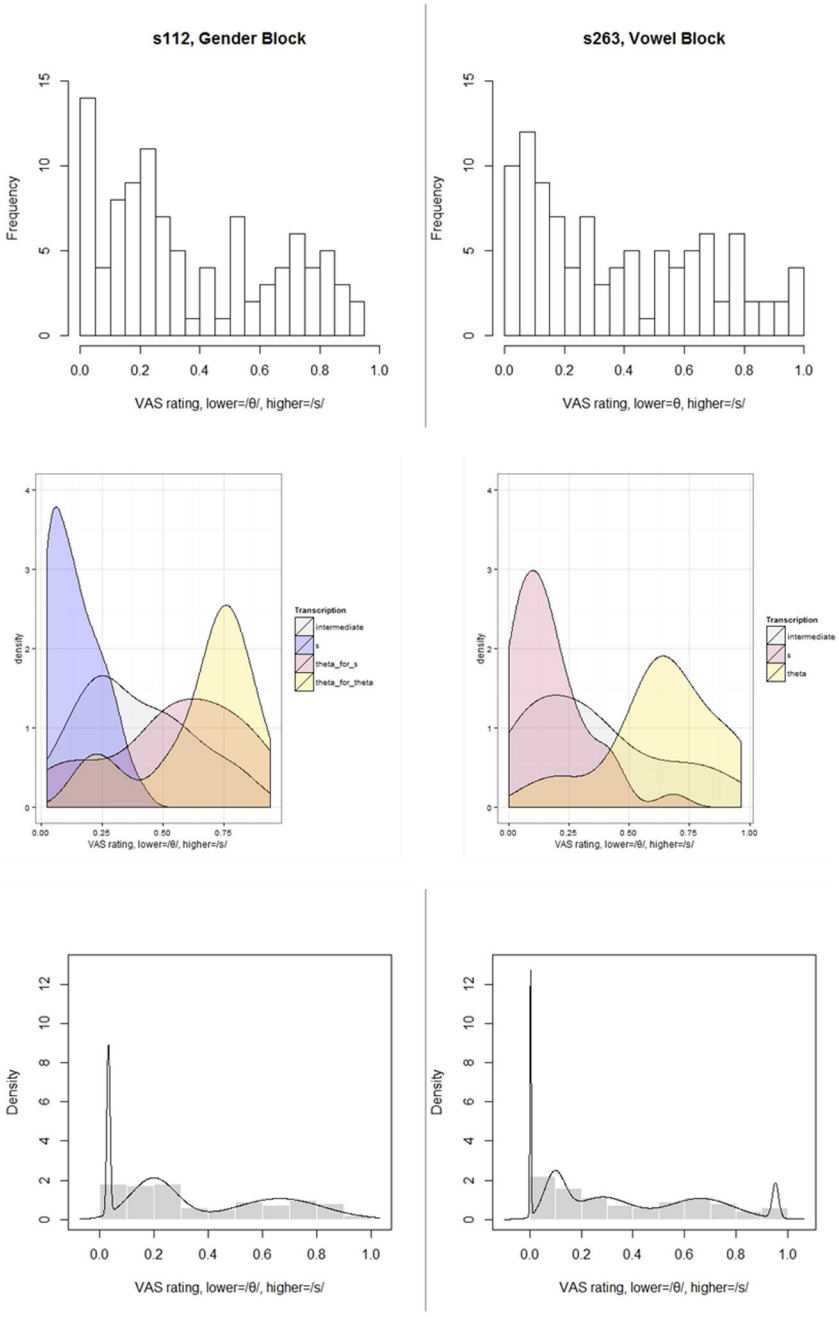

Figure 5.
Examples of analyses of the gradiency of response used in Experiment 1.

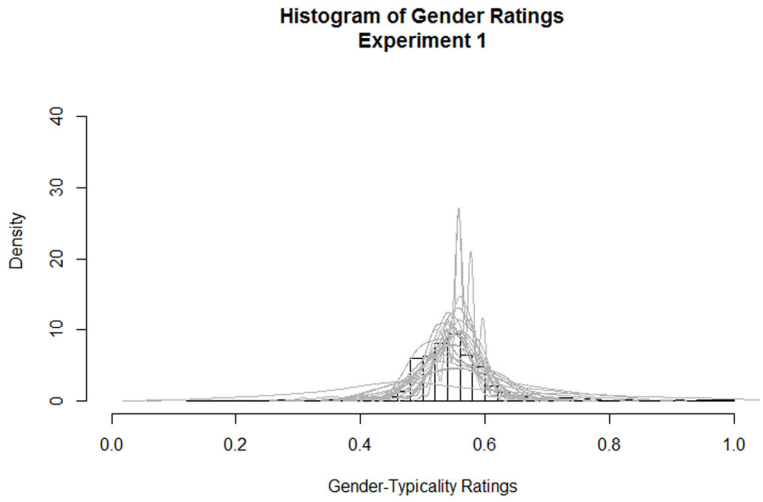

Figure 6.
Distribution of individual listeners' gender ratings in Experiment 1.

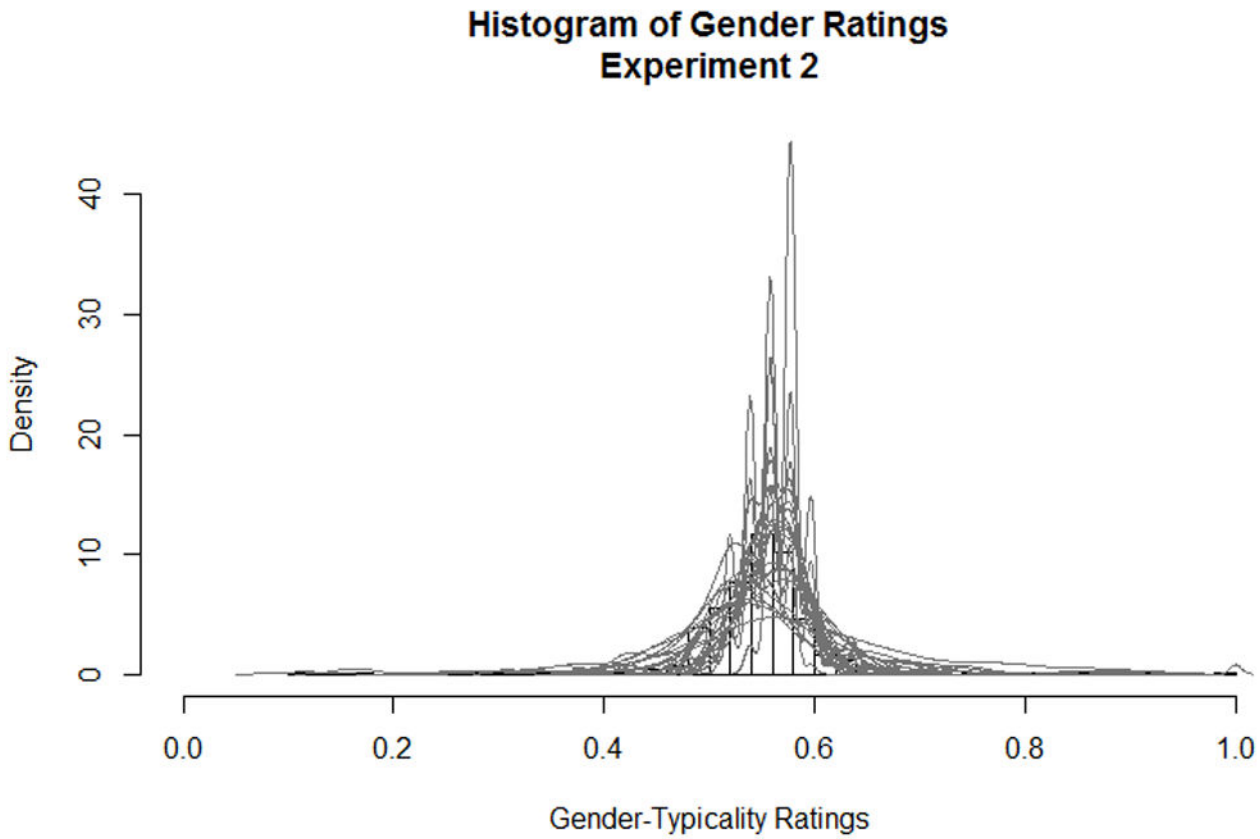

Figure 7.
Distribution of individual listeners' gender ratings in Experiment 2.

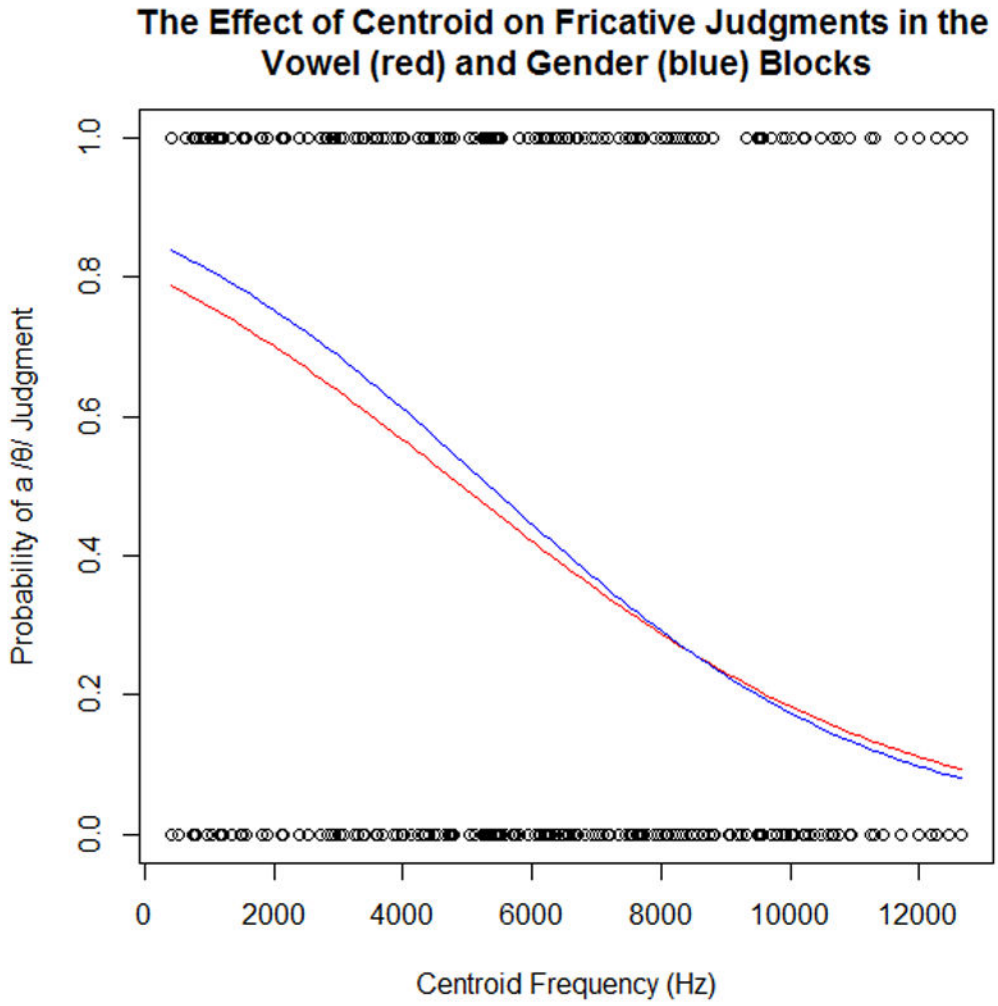

Figure 8.
Logistic functions for Experiment 2, predicting fricative judgments from fricative M1, separated by condition.

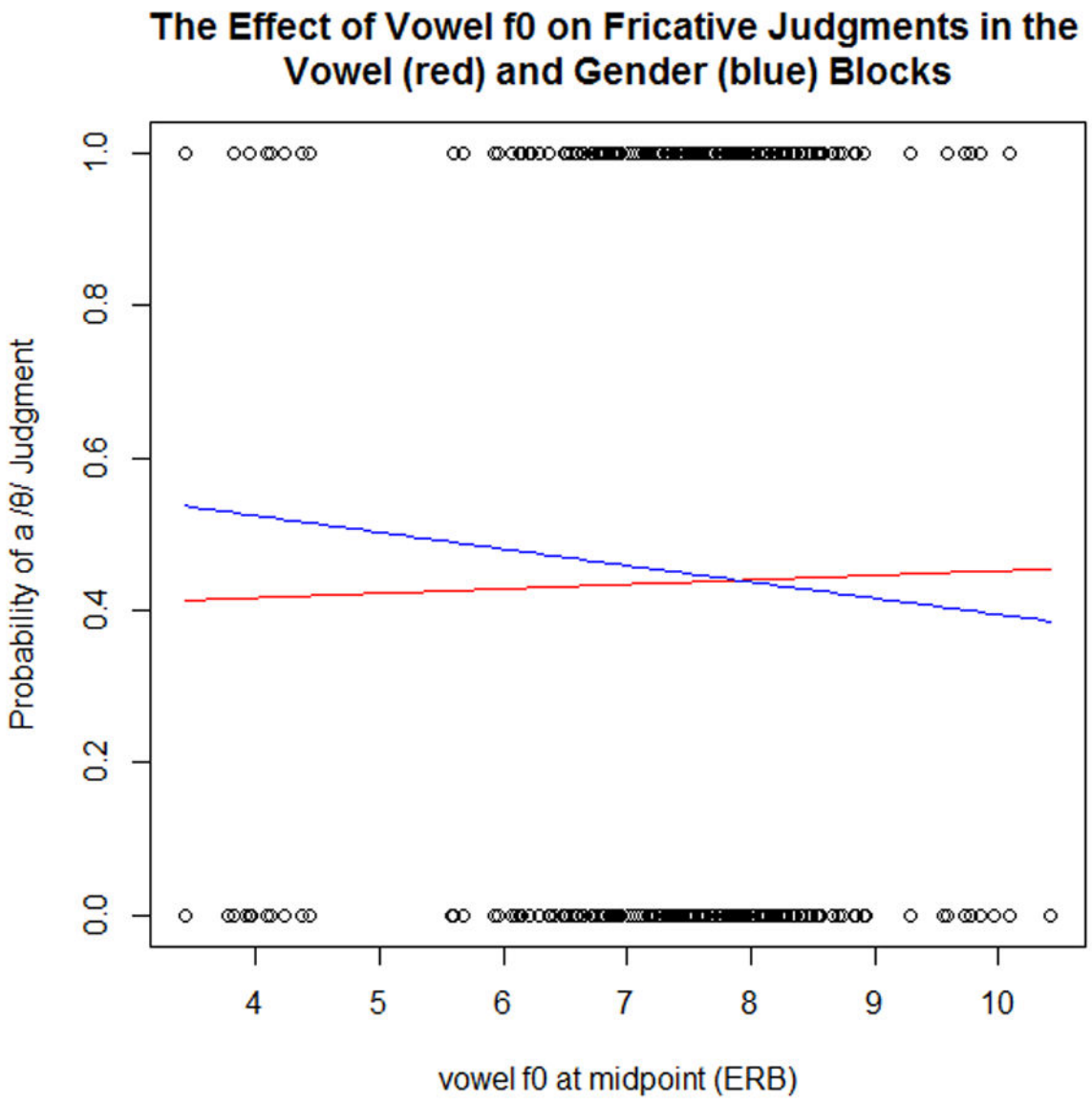

Figure 9.
Logistic functions for Experiment 2, predicting fricative judgments from vowel f0, separated by condition.

Table 1

Mean (and Standard Deviation) for selected acoustic measures of the consonant and vowel portions of the stimuli, separated by transcription categories

Measure	[s] for /s/ (n=50)	[s] for /θ/ (n=26)	[s:θ] (n=24)	[θ:s] (n=30)	[θ] for /s/ (n=24)	[θ] for /θ/ (n=46)
M1 (Hz)[a]	83841833	5667(3103)	6223(2165)	6285(2604)	54912884	3338(3621)
M2 (Hz)[b]	2807 (917)	3310 (941)	4194 (927)	36091087	3924(1236)	3742(1351)
Onset F2 (Bark)[c]	13.82(1.04)	13.92(1.87)	13.90(1.01)	13.65(1.52)	13.51(1.13)	13.83(1.35)
Midpoint f0 (ERB)[d]	7.40 (1.14)	7.21 (1.49)	7.46 (1.12)	7.53 (1.06)	7.10 (1.14)	7.30 (1.30)
Fricative Duration (ms)[e]	124 (52)	80 (26)	128 (40)	132 (49)	138 (56)	94 (30)
Relative Intensity (dB)[f]	−12.4 (5.3)	−13.8 (6.4)	−17.2 (4.7)	−17.8 (4.5)	−18.7 (5.6)	−23.5 (5.4)

[a] First spectral moment (centroid) of a 40 ms interval of frication noise at midpoint,

[b] Second spectral moment (spectral spread) of a 40 ms interval of frication noise at midpoint,

[c] Second formant frequency of the vowel at its onset, in Bark units,

[d] Fundamental Frequency of the vowel at midpoint,

[e] Duration of the fricative,

[f] Difference in RMS intensity of the vowel and the fricative

Table 2

Linear mixed-effects model predicting continuous ratings of fricative goodness in Experiment 1 from selected acoustic measures of the stimuli.

Variable	Estimate	Std. Error	df	t-value	Pr(>|t|)
(Intercept)	0.4164	0.0203	21.0	20.49	<0.001
M1[a]	−0.0687	0.0062	26.0	−11.06	<0.001
M2[b]	0.0285	0.0048	21.0	5.99	<0.001
onset F2[c]	−0.0077	0.0042	21.0	−1.81	0.085
Midpoint f0[d]	−0.0216	0.0041	23.0	−5.29	<0.001
Duration[e]	0.0338	0.0040	40.7	8.53	<0.001
Relative Intensity[f]	−0.1084	0.0100	22.0	−10.88	<0.001

All predictor variables were z-transformed prior to analysis.

[a]First spectral moment (centroid) of a 40 ms interval of frication noise at midpoint,

[b]Second spectral moment (spectral spread) of a 40 ms interval of frication noise at midpoint,

[c]Second formant frequency of the vowel at its onset, in Bark units,

[d]Fundamental Frequency of the vowel at midpoint,

[e]Duration of the fricative,

[f]Difference in RMS intensity of the vowel and the fricative.

Table 3

Logit mixed-effects model predicting categorical judgments of fricative type in Experiment 2 from selected acoustic characteristics of the stimuli, and from condition (i.e,. whether they were made in the vowel block or the gender block).

Variable	Estimate	Std. Error	z-value	Pr(>|z|)
(Intercept)	0.4517	0.1046	−4.32	<0.001
M1[a]	−0.9853	0.0539	−18.27	<0.001
M2[b]	0.5986	0.0469	12.76	<0.001
onset F2[c]	−0.1507	0.0389	−3.87	<0.001
Midpoint f0[d]	−0.3403	0.0406	−8.37	<0.001
Duration[e]	0.3738	0.0456	8.20	<0.001
Relative Intensity[f]	−1.3129	0.0635	−20.69	<0.001
M1 * Condition[g]	0.1430	0.0485	2.95	0.003
Midpoint f0 * Condition[g]	0.0934	0.0403	2.32	0.020

All predictor variables were z-transformed prior to analysis.

[a]First spectral moment (centroid) of a 40 ms interval of frication noise at midpoint,

[b]Second spectral moment (spectral spread) of a 40 ms interval of frication noise at midpoint,

[c]Second formant frequency of the vowel at its onset, in Bark units,

[d]Fundamental Frequency of the vowel at midpoint,

[e]Duration of the fricative,

[f]Difference in RMS intensity of the vowel and the fricative,

[g]Condition was contrast coded: −1=fricative judgments made in the vowel block, 1=fricative judgments made in the gender block.

Table 4

Logit mixed-effects model predicting categorical judgments of fricative type in Experiment 1 (derived by converting the continuous ratings into categorical ones) from selected acoustic characteristics of the stimuli.

Variable	Estimate	Std. Error	z-value	Pr(>|z|)
(Intercept)	−0.6781	0.1808	−3.75	<0.001
M1[a]	−0.5819	0.0608	−9.56	<0.001
M2[b]	0.3011	0.0521	5.78	<0.001
onset F2[c]	−0.0748	0.0410	−1.82	0.068
Midpoint f0[d]	−0.2798	0.0414	−6.75	<0.001
Duration[e]	0.2939	0.0429	6.85	<0.001
Relative Intensity[f]	−1.1444	0.0950	−12.05	<0.001

All predictor variables were z-transformed prior to analysis.

[a]First spectral moment (centroid) of a 40 ms interval of frication noise at midpoint,

[b]Second spectral moment (spectral spread) of a 40 ms interval of frication noise at midpoint,

[c]Second formant frequency of the vowel at its onset, in Bark units,

[d]Fundamental Frequency of the vowel at midpoint,

[e]Duration of the fricative,

[f]Difference in RMS intensity of the vowel and the fricative.