



Published in final edited form as:

Clin Linguist Phon. 2010 January ; 24(4-5): 245–260. doi:10.3109/02699200903532524.

Deconstructing Phonetic Transcription: Covert Contrast, Perceptual Bias, and an Extraterrestrial View of *Vox Humana*

Benjamin Munson^{1,a}, Jan Edwards², Sarah Schellinger^{1,2}, Mary E. Beckman³, and Marie K. Meyer¹

¹Department of Speech-Language-Hearing Sciences, University of Minnesota, Minneapolis

²Department of Communicative Disorders, University of Wisconsin, Madison

³Department of Linguistics, Ohio State University, Columbus

Abstract

This article honours Adele Miccio's life work by reflecting on the utility of phonetic transcription. The first section reviews the literature on cases where children whose speech appears to neutralize a contrast in the adult language are found on closer examination to produce a contrast (*covert contrast*). We present evidence from a new series of perception studies that covert contrast may be far more prevalent in children's speech than existing studies would suggest. The second section presents the results of a new study designed to examine whether naïve listeners' perception of children's /s/ and /θ/ productions can be changed experimentally when they are led to believe that the children who produced the sounds were older or younger. Here, it is shown that, under the right circumstances, adults report more tokens of /θ/ to be accurate productions of /s/ when they believe a talker to be an older child than when they believe the talker to be younger. This finding suggests that auditory information alone cannot be the sole basis for judging the accuracy of a sound. The final section presents recommendations for supplementing phonetic transcription with other measures, to gain a fuller picture of children's production abilities.

Keywords

phonetic transcription; speech perception; covert contrast; phonological acquisition

One Memory of Adele Miccio: An Inspired Rant about Voiceless Lateral Fricatives

Adele Miccio shared at least two characteristics in common with the first two authors of this article. The first is that all of our research integrates knowledge and methods from speech-language pathology with those from linguistics. The second is that all three of us have taught undergraduate students phonetic transcription. It was in a discussion about those two facts that the first author had a very passionate exchange with Adele at an American Speech-Language-Hearing Association annual meeting about the proper transcription of misarticulations of /s/ with high lateral airflow (i.e. so-called laterally lisped /s/). The specific question that we were debating was whether such productions should be transcribed with an [s] symbol and a diacritic indicating lateral airflow, or whether we should simply use the existing phonetic symbol for this sound when it occurs in normal speakers in languages like Zulu or Welsh, the voiceless

^aContact Author: Department of Speech-Language-Hearing Sciences, 115 Shevlin Hall, 164 Pillsbury Drive, SE, Minneapolis, MN 55455, munso005@umn.edu, +1 612 624 0304, Fax +1 612 624 7586.

lateral fricative, [ɬ]. This argument was particularly memorable because of the contrast between its surface absurdity (how could two people discuss so passionately and for so long something as seemingly trivial as the proper way of transcribing a sound?) and the deeper topics that it touched on (what is the relationship between phonetic variation and the symbols that we use to note it?).

This article memorialises Adele Miccio by discussing phonetic transcription. It is a philosophical think-piece, a review of some of our recent research on this topic, and a report of a new set of experiments designed to examine how people perceive children's speech. At first glance, this might seem akin to memorialising Senator Edward Kennedy with an essay on parliamentary procedure. But just as many important pieces of legislation live (and die) because of the intricacies of parliamentary procedure, so does much of our knowledge of spoken language rest on the process of phonetic transcription. We can think of no better way to remember Adele Miccio than to encourage people to think about the very foundation of our understanding of spoken language.

Human speech: The extraterrestrial view

As researchers and clinicians, we phonetically transcribe speech nearly every day. The practice of phonetic transcription is so entrenched in our lives and in the study of human language that it is difficult to deconstruct it in order to evaluate the component assumptions on which it is based. To help the reader do this, imagine the following scenario. A group of peaceful extraterrestrial beings arrive on Earth. These creatures communicate solely in the thermal modality, using a set of organs that have evolved to generate and sense rapid temperature fluctuation patterns. The aliens have come to Earth as part of a large project, funded by their enlightened alien government, to describe variation in life throughout the travelled universe. Understandably, the aliens would be interested in describing the animals living on Earth. In describing the higher primates, they would undoubtedly note that one primate species, *homo sapiens sapiens*, differs from the other species in (among other things) its use of a complex symbolic communication system.

Describing this system would be a daunting task. We might imagine that after they have grasped the difference in modality, the aliens would use the same tactic taken by many humans when studying an unfamiliar language, and begin by describing the sound system of the Earthling language. We can expect that the scientific progress that allowed these aliens to travel to Earth would also have resulted in them being expert comparative anatomists, physiologists, and acousticians. Hence, the aliens would be able to describe speech-sound production and its acoustic consequences. First of all the aliens would note that humans use a small set of anatomical structures and articulatory manoeuvres to produce sounds: air is forced out of or drawn in to the oral cavity, the nasal cavity, or both. Different sound qualities result from contorting these cavities (through movements of the tongue, the velum, and the lips) so that they have different shapes and different degrees of stricture, and by manipulating air pressure changes during production. As the aliens continued in their linguistic fieldwork, they also would have the opportunity to examine the task of speech acquisition. Here the aliens would no doubt note that children do not achieve fully adult-like speech until relatively late in development, especially compared to other complex motor tasks such as locomotion or reaching for an object.

What is not clear, however, is whether these fictional alien anthropologists would come up with anything remotely like phonetic transcription (such as the International Phonetic Alphabet [IPA]) to characterize human speech. That is, it is not inevitable that the aliens would use the symbol [s] (or some other arbitrary symbol) to denote both the first sound in the Japanese word 寿司 and the English borrowing *sushi*, nor would they use the symbol [ʃ] to denote the sound

at the beginning of the second syllable in that word. They would likely not use the symbol [s] to denote the misarticulations that human speech-language pathologists have come to call depalatalisation errors (such as productions of *shoe* that sound like *sue*).

The remainder of this article is to describe why this is so. The first section describes the limitations in the denotational system that arise because of its categorical nature.

Covert Contrast is Everywhere

By its very nature, the IPA is a categorical system. A fixed number of categories—symbols and diacritics—are used to denote continuous variation in speech sounds. One problem that arises because of this occurs when we observe speech-sound variation with finer-grained observational tools. Such investigations often result in the observation that speech sound development is not necessarily categorical; children's productions do not always progress directly and categorically from incorrect to correct. Before children produce a contrast between two sounds, they may produce a 'covert contrast,' a subphonemic difference that is typically not large enough to warrant being transcribed by a different phonetic symbol, but can be measured acoustically. Covert contrast was first robustly documented in the literature by Macken and Barton (1980), for the voicing contrast in stops. In a longitudinal study of four children, they observed that these children went through a phase where most of their productions of voiceless stops were perceived as voiced, even though the children were producing longer VOTs on average for the target voiceless stops relative to the target voiced stops. The impression of systematic substitution of voiced for voiceless stops was because all of the productions had VOTs that were in the adult voiced range for English. That is, covert contrasts might lead children to be identified as having an 'initial obstruent voicing' error pattern when in fact they were producing voicing distinctions. Since this seminal paper, many researchers have found acoustic evidence of covert contrast in the speech of both children with typically developing production skills and children with phonological disorders. Covert contrast has been observed for a variety of contrasts, including place of articulation for stops (e.g., Forrest et al., 1990), place of articulation for fricatives (e.g., Baum and McNutt, 1990; Li et al., 2009), and voicing for stops (e.g., Macken and Barton, 1980; Maxwell & Weismer, 1982). Covert contrast is also clinically important; Tyler and colleagues (Tyler et al., 1993) found that children who exhibited a covert contrast made more rapid progress in therapy than children who exhibited no contrast at all. Even when it is not documented acoustically, studies of intra-child variability in production strongly suggest the presence of covert contrast, as shown in Hewlett and Waters' (2004) review of phonological development studies.

This research on covert contrast has had relatively little influence on clinical practice. At least one reason for this is because clinically feasible methods of acoustic analysis have not yet been developed. A second is that studies that have used acoustic analyses have found covert contrasts in relatively few children. We suspect that the reason that only a few cases of covert contrast are evidenced in acoustic studies is because of the nature of acoustic analysis. While the acoustic signal itself is rich and redundant, acoustic analyses typically focus on only a few specific parameters in order to study phonetic contrasts. Part of this is likely for the sake of expediency, but part is based perhaps on the mistaken belief that phonemes or features have a single invariant acoustic correlate. For example, as discussed above, VOT is the primary cue to the voicing contrast for stop consonants, and studies that have looked for a covert voicing contrast have focused on VOT. However, there are a number of other cues to stop voicing besides VOT even in utterance-initial position where closure duration and preceding vowel duration cannot be a cue – for example, fundamental frequency at the onset of the following vowel (Haggard, Summerfield, & Rogers, 1981), the amplitude of aspiration relative to that of the following voiced part of the vowel (Repp, 1979), and differences in the ratio of the first harmonic to the second harmonic also serve to cue the contrast between voiced and voiceless

stop consonants in English (Kong, 2009). Moreover, the relative importance that these different features play differs across languages. Kong (2009) showed that aspiration intensity plays a greater role in differentiating among voicing categories in Korean than in English. However, since voice onset time in and of itself is adequate to distinguish between voiced and voiceless stops in initial position in adult productions in English, few researchers have looked for evidence of covert contrast for voicing in other parameters in this position. It may be that we see relatively little instance of covert contrast in acoustic analyses of production because of the reductionist nature of acoustic analysis; that is, we look at only a few cues and we examine these cues separately.

The results of a series of perception experiments that we have conducted over the past several years support this interpretation of the spotty evidence for covert contrast to date (e.g. Schellinger, Edwards, Munson, and Beckman, 2008; Urberg-Carlson, Kaiser, and Munson, 2008). More generally, these results suggest that covert contrast in acquisition is the rule rather than the exception. These experiments were originally designed to examine the relationship between perception of particular contrasts by naïve listeners and the acoustic parameters that differentiate these contrasts. The stimuli for these experiments came from the παιδολογος ([paidologos]) data base described in Edwards and Beckman (2008). The word-initial consonants in this data base were transcribed by an adult native speaker using four categories: correct (e.g. [t] for /t/), clear substitution ([k] for /t/), intermediate between two sounds ([t]:[k] means ‘in between /t/ and /k/ but more like /t/’, while [k]:[t] means ‘in between /t/ and /k/ but more like /k/’), and distortion (such as a lateral lisp – a production that is not possible to transcribe with conventional IPA, although the version of the IPA with extensions for disordered speech [extIPA, Ball & Müller, 2005] does have a wider range of symbols for different types of speech distortions). The perception experiments included all of the transcription categories except distortions. For example, the perception experiment on the contrast between /s/ and /θ/ included correct /s/ productions, correct /θ/ productions, [θ] for /s/ substitutions, [s] for /θ/ substitutions, and productions intermediate between /s/ and /θ/ (both [s]:[θ] and [θ]:[s]). Other contrasts that have been studied include the contrast between alveolar and velar stop consonants, the contrast between /s/ and /ʃ/, and the contrast between voiced and voiceless stop consonants. The method used in the perception experiments was visual analog scaling or VAS (Urberg-Carlson et al., 2008). In VAS rating tasks, individuals are asked to scale a psychophysical parameter by indicating their percept on an idealized visual display. In the VAS tasks reported by Schellinger et al. and Urberg-Carlson et al., listeners were presented with a horizontal line with an orthographic label of each of the two sounds as endpoints (for example, ‘s’ as the label for /s/ would be at one endpoint and ‘th’ as the label for /θ/ would be at the other, with clear instructions that ‘th’ should be interpreted as the voiceless variant) and are asked to click on the line location that represents where each production falls on the continuum between /s/ and /θ/. For the two experiments discussed in this section, the listeners were 20 adult native speakers of English. For the /s/-/θ/ contrast, all of the stimuli were word-initial consonant-vowel (CV) sequences excised from words produced by English speakers. For the /d/-/g/ contrast, the stimuli included word-initial /d/-/g/ produced by English speakers. Listeners in /s/-/θ/ experiment were native speakers of English. Listeners in the other experiment were either native speakers of English or native speakers of Greek, as this experiment was done as part of a larger project examining the relative contribution of speaker- and listener-related factors on the acquisition of phonology.

Figure 1 below shows the results for the /s/-/θ/ and the /d/-/g/ contrasts (taken, respectively, from Schellinger et al., 2008 and work in progress by Arbisí-Kelm, Edwards, and Munson). The same pattern is observed in both figures. In Figure 1, all of the transcription categories are significantly different from each other. In Figure 2, all of the transcription categories except for [d] for /g/ substitutions versus correct /d/ productions are significantly different from each other for the English-speaking listeners. The same pattern was observed for the other two

contrasts that we have examined, that between /t/ and /d/ and that between /s/ and /ʃ/ (Kong, 2009; Urberg-Carlson et al., 2008, respectively). Figure 2 also illustrates that these ratings show strong effects of listener language. The Greek speakers, for example, rated the English velar tokens as more-front than the English-speaking listeners did. This cross-language asymmetry indicates that this perception is mediated by language-specificity in the way that phonological contrasts are expressed, a topic that we explore further in Arbisi-Kelm, Beckman, Kong, and Edwards (2008).

While we were not surprised that naïve listeners could distinguish between correct and intermediate productions, we were somewhat surprised that they consistently distinguished between correct productions and clear substitutions. That is, naïve listeners consistently perceived differences between [d] for /g/ substitutions, and correct /d/ productions, between [θ] for /s/ substitutions and correct /θ/ productions, between [s] for /ʃ/ substitutions and correct /s/ productions, and between [d] for /t/ substitutions and correct /d/ productions. In all of these cases, the substitution was judged as less target-like than the correct production. We hasten to note that ours are not the only studies that have found evidence that listeners perceive consonants gradually. As part of their critique of phonetic transcription as a tool in sociolinguistic research, Kerswill and Wright (1990) show that listeners report different proportions of ‘d’ precepts in stimuli taken from d#g sequences with varying degrees of overlap between the alveolar and dorsal gestures in the /d/ and /g/.

These results suggest that covert contrast is ubiquitous. It may well be the rule, rather than the exception. We suspect that it is easier to find evidence of covert contrast in a perception task than in an acoustic analysis because listeners are presented with the richness of the entire acoustic signal, while acoustic analysis focuses on only one parameter or, at best, a few parameters. We want to make clear that we are not suggesting that categorical perception does not exist. Rather, we argue, as have others, that categorical perception is a consequence of the task used to measure perception: whether a listener perceives a sound as categorical or not depends on the extent to which the task requires strict categorization. When the task promotes the perception of categories (either because of the difficulty of the task itself, or because of the use of categorical labels), people behave as if they can only hear categories and not the phonetic detail that these categories subsume. When different methods are used, individuals show exquisite sensitivity to the phonetic variation within categories. When the trained native speaker/transcriber was asked to place the [d] for /g/ productions or the [θ] for /s/ productions into a category, she labelled them as clear substitutions – not as intermediate productions or distortions. But when naïve listeners were asked to rate these same productions on a continuum, they heard them as less target-like than productions that has been transcribed as correct.

Perceptual Bias

Imagine again our alien anthropologists. As they continued their study of the sound structure of languages, they would surely note that there is considerable variation within a language in the articulatory and acoustic characteristics of speech sounds, and that some of these differences can be predicted by attributes of speakers. They might note (as did Langstrof, 2006) that there is considerable variation in New Zealand in the pronunciation of the vowels in the words *trap* and *dress*, such that older speakers' productions of the vowel in *dress* resemble younger speakers' production of the vowel in *trap*. They would also likely note that many listeners in these dialects are able to understand speech despite these sometimes stark variations among groups of talkers. That is, many listeners appear to have a rich enough knowledge of how sounds vary across social groups that they are able to parse out this variability when perceiving speech.

An emerging body of literature has demonstrated experimentally how readily listeners calibrate their perception when led to expect a talker to produce a particular variant of a sound. Drager (in press), for example, showed that listeners in New Zealand calibrate their expectations about vowel productions based on presumptions regarding a speakers' age. The speakers' apparent ages were manipulated by pairing speech tokens with pictures of either an older adult or a younger adult. The direction of the effect was exactly as predicted by Langstrof's production data: vowels that were acoustically intermediate between those in *dress* and *trap* were more likely to be identified as *trap* when the listeners believed they were produced by a younger speaker, and as *dress* when produced by an older one.

These findings have clear implications for the topic of this article, the perception of children's speech. Whether we are talking about phonetic transcription or about other types of rating, like VAS, we would like to know what listeners' responses reflect. Ideally, they reflect the articulatory and acoustic characteristics of the sound being transcribed or rated. We cannot rule out, however, that adults' perception of children's speech is similarly affected by social biases, just as their perception of other adults' speech is. Indeed, this conjecture is made all the more plausible by the existence of many social stereotypes about how children speak. For example, the stereotype in English-speaking cultures that young children substitute [t] and [d] for /k/ and /g/ is encapsulated in Dorothy Parker's report that 'Tonstant Weader fwowed up' (in her 1928 review of A. A. Milne's *The house at Pooh corner*), as well as in Samuel Butler's description (in his 1903 autobiographical novel *The way of all flesh*) of being punished for making this substitution. Similarly, the stereotype that young children substitute [s] for /ʃ/ is at least as old as Elizabeth Gaskell's last novel *Wives and daughters* (published after her death in 1865), which includes a passage where a toddler is transcribed as saying *I s'ant* for *I shan't*.

Given these cultural stereotypes, we might wonder whether children's intermediate productions, such as those described in the previous section, are particularly susceptible to bias about the age of speakers. That is, when listeners are presented with something that isn't a clear endpoint, are they likely to rate it differently depending on whether they think the speaker is a younger child or an older one?

In this section we report on an experiment (with three conditions) designed to examine this possibility. The experiment is a follow-up to the experiment presented by Schellinger, Edwards, Munson, and Beckman (2008). In that experiment, Schellinger et al. examined adults' perception of 200 tokens of children's productions of target /s/ and /θ/, taken from the *παιδολογος* database of children's speech (Edwards and Beckman, 2008). The stimuli were sets of approximately equal numbers of productions in six categories, as described earlier. Recall that Schellinger et al. conducted a VAS experiment and confirmed that naïve listeners rated all six of these fricative types differently from one another.

Schellinger et al. also conducted a second experiment in which they played listeners these sounds preceded by carrier phrases. One of these carrier phrases was a recording of a young child saying 'I really like.' The other was a recording of the same child saying 'I weawwy yike,' i.e. saying the same phrase but with stereotypical developmental speech-sound errors targeting /r/ and /l/. Multiple recordings of each carrier phrase type were used. Half of the 'really like' carrier phrases were acoustically modified so that the formant frequencies and fundamental frequency were lower than in the natural recording, consistent with the productions of an older child. Half of the 'weawwy yike' recordings were scaled in the opposite direction, consistent with the productions of a younger child. Pre-testing with an independent group of listeners showed that the talker of the 'weawwy yike' carrier phrases was consistently perceived to be younger than the talker of the 'really like' carrier phrases, regardless of whether the carrier phrases had been scaled acoustically. Hence, both modified and unmodified carrier

phrases were mixed within a block to increase the number of acoustically distinct carrier phrases and thus to decrease the likelihood that listeners would realize that many of them were identical. ‘Really like’ and ‘weawwy yike’ carrier phrases were presented in a single block of 400 tokens (i.e. each of the 200 tokens was presented in two different trials, once preceded by a ‘really like’ and once by a ‘weawwy yike’ carrier phrase) in fully random order.

In the perception task, Schellinger played a carrier phrase followed by a token, and asked listeners to judge whether it was an acceptable token of the sound ‘s’. The proportion of ‘yes’ responses was calculated separately for each of the six fricative types preceded by ‘really like’ and ‘weawwy yike’ carrier phrases. As with the VAS task, the proportion of ‘yes’ responses differed for each of the six fricative types. However, only a small biasing effect of carrier-phrase type was found. The current experiment follows up on this finding.

As noted earlier, the current experiment has three conditions. The first condition examined whether stronger biasing could be obtained by blocking the perception task by carrier-phrase type. We reasoned that blocking by carrier phrase would encourage the listeners to more consistently calibrate their criteria for an acceptable token of /s/.

The second condition examined whether the perception of /s/ can be affected by the instructions that listeners are given in the perception task. In both Schellinger et al. and in condition 1 listeners were told that the purpose of the project was to examine the perception of developmental misarticulations of /s/. This explicit mention of ‘misarticulation’ might have led the listeners to respond qualitatively differently from how they would have responded if ‘misarticulation’ had not been mentioned. Condition 2 tested this by examining the performance of listeners in a task that was blocked by carrier phrase type (as with Condition 1), but which did not mention developmental misarticulations in the instructions.

Condition 3 examined whether greater biasing could be obtained when carrier phrases were acoustically modified to resemble the target fricative-vowel stimuli acoustically. Here we reasoned that acoustically matching the carrier phrase and the target would increase the likelihood that the listeners would be willing to imagine them as being produced by the same talker. The greater acoustic similarity was achieved by matching the peak f0 of the carrier phrase with the average f0 of the vowel in the stimulus. Table I summarizes the different experimental conditions.

Methods

Subjects

Fifteen listeners participated in each of the three conditions. The listeners were recruited from the University of Minnesota community through fliers on campus. They included a mix of undergraduate students, university staff, and visitors to the university. The average age for participants in Condition 1, 2, and 3 was 22.5 (SD = 5.1), 23.9 (SD = 8.1), and 25.1 (SD = 9.6) respectively. The listeners had limited experience with hearing children's speech, as measured by self ratings. They were asked, on a scale from 1-10, how much time they spent around children under the age of 5 years, with 1 being no time at all and 10 being most of their time. The average ratings for participants in Condition 1, 2, and 3 were 2.2 (SD = 1.9), 2.9 (SD = 2.5), and 3.7 (SD = 2.5) respectively. None of these differences was significant in a Kruskal-Willis nonparametric test.

Stimuli

The stimuli were 200 fricatives taken from the *παιδολογος* database. They were produced by 2- through 5-year-old children acquiring English monolingually, and were elicited through real-word and nonword repetition tasks in which children saw a picture of a familiar object (in

the real word task) or a novel object (for the nonword task) and heard an accompanying production of the word or nonword. They then repeated the audio prompt. Children's productions were transcribed by two experienced native-speaker transcribers who were unaware of what the target consonant was.

The stimuli were analysed acoustically. The results of this analysis are presented in Table II. Briefly, a spectrum was calculated over the middle 40 ms of each fricative, to derive three spectral measures: the fricative's overall loudness, its peak frequency, and a measure of the distribution of energy around the peak (the 'compactness index'). Measures were based on psychophysically transformed spectra (i.e. examining loudness in sones rather than intensity in decibels, and frequency in equivalent rectangular bandwidths [ERB] instead of hertz). Additionally, measures of duration and of the second-formant frequency at vowel onset (in ERB) are reported in Table II. A defence of the psychophysical measures, as well as an illustration of their benefit over traditional linear measures, can be found in Arbisi-Kelm, Beckman, Kong, and Edwards (2008).

The carrier phrases were the same as in Schellinger et al. (2008), described earlier. For condition 3, the fundamental frequency of the carrier phrase was scaled using the PSOLA algorithm in Praat (Boersma and Weenink, 2009), such that the f_0 of the carrier phrase at its offset was equal to the average f_0 of the vowel portion of the target CV. This scaling was chosen in a pre-test in which a group of listeners who did not participate in any other experiment was played a set of 10 stimuli preceded by carrier phrases that were scaled to different f_0 s relative to the target CV and were asked to choose the pairs of stimuli that sounded most like they were produced by the same child. The pairs whose carrier phrase offset f_0 s were identical to the average f_0 of the CV were most often chosen as the best match.

Procedures

All three tasks were administered with the E-Prime experiment design and management software. Participants in Conditions 1 and 3 were given instructions that mentioned 'speech-sound delays or disorders.' Specifically, they were told that they 'may hear 's' productions incorrectly produced as 'th,' due to a common error called a frontal lisp.' Participants in Condition 2 were given instructions that made no mention of a lisp or speech-delays or disorders. Participants in all three conditions were instructed to expect to hear the phrase, 'I really like' followed by a consonant-vowel sequence starting with 's.' When asked 'Is the 's' sound correct?' they were to respond 'yes' or 'no' using a button box whose buttons were labelled clearly.

Analysis

For each condition, the proportion of 'yes' responses for each of the six fricative types was calculated separately for each of the two carrier phrases. These proportions were submitted to a three factor (6 fricative type \times 2 carrier phrase \times 3 condition) within-subjects Analysis of Variance. Effect sizes were calculated for each significant factor. Bonferroni-corrected post-hoc paired comparisons were used to compare differences among fricative types.

Results

Figure 3 shows the proportion of 'yes' responses in the two carrier phrases for Condition 1 (Figure 3, left), Condition 2 (Figure 3, center), and Condition 3 (Figure 3, right). The effect of transcribed fricative type was both statistically significant and very large, $F[5,210] = 247.7$, $p < 0.001$, $\eta^2_{\text{partial}} = 0.86$. Post-hoc Bonferroni-corrected paired comparisons showed significant differences between all pairs of fricative types, in the direction that would be predicted based

on the VAS ratings reported by Schellinger et al. (2008). The effect of carrier phrase type was also significant, though its effect was considerably smaller than the effect of fricative, $F[1,42] = 4.6$, $p = 0.038$, $\eta^2_{\text{partial}} = 0.10$. Sounds preceded by ‘really like’ carrier phrase were more likely to be judged as correct /s/ than those preceded by the ‘weawwy yike’ carrier phrase. The effect of condition was also significant, $F[1,42] = 4.2$, $p = 0.021$, $\eta^2_{\text{partial}} = 0.17$. Post-hoc tests showed that more ‘yes’ responses were given in condition 2 than in condition 1. Neither of the other two comparisons showed statistically significant differences.

Finally, there was a two-way interaction between condition and fricative type, $F[4.4,92.6] = 4.3$, $p = 0.002$, $\eta^2_{\text{partial}} = 0.17$. This interaction occurred because the overall higher rates of ‘yes’ responses in condition 2 affected the more /θ/-like sounds more than it affected the /s/-like sounds, whose ratings were close to ceiling. Hence, there was a larger effect of condition on /θ/-like sounds than on /s/-like sounds. This finding was quite unexpected, and likely relates to the unique status of frontal errors for /s/. There exist in North America and elsewhere popular-culture associations between frontal /s/ and different social categories. As shown by Munson and Zimmerman (2006), listeners label male talkers as less prototypically heterosexual sounding when their speech contains frontal /s/. Moreover, there is considerable variation within and across languages in the tendency to produce frontal variants of /s/. As shown by Dart (1991), women are more likely to produce more-frontal variants of this sound than men, and French speakers of both sexes produce a more-frontal /s/ than English speakers. Listeners simply expect that /s/ variation is part of normal phonetic variation in adults’ speech. Hence, when listeners were not told that the study related to developmental *mis*articulations, they were more willing to interpret the /θ/-like tokens as variants of /s/ than when they were told explicitly that they were participating in a study on misarticulation.

This explanation might help explain some of the other response patterns that we observed. Consider first Figure 5. This figure shows that carrier phrase type had a larger influence on ratings of the /θ/-like stimuli than on ratings of the /s/-like ones. They were less likely to be treated as errors of /s/ when preceded by the ‘really like’ carrier phrases than when preceded by the ‘weawwy yike’ ones. One interpretation of this difference is that when listeners thought they were listening to an older child, they treated /θ/-like pronunciations as normal variation in target /s/, of the type you might expect to observe in adults. When they thought they were listening to a younger child, they treated these as target /θ/. Interestingly, this pattern was not seen in condition 1, which differed from condition 3 only in that it didn’t match the f_0 of the carrier phrase to the f_0 of the targets. Figure 3 shows that adults in condition 1 were biased more on the /s/-like stimuli. If the f_0 -matching of condition 3 had the intended effect of allowing the listeners to interpret the carrier phrase and the target as having been produced by the same child, then we imagine that the results in that condition are a more-faithful representation of the kind of biasing that would exist in real-world listening tasks.

This effect seen in condition 3 is rather surprising, and is the direct opposite of what we would predict based on other studies that we have done recently. Munson (2009) examined the perception of an /s/-/θ/ continuum combined with vocalic bases (to create a series of *sigh-thigh* continua). Some of the vocalic bases were acoustically altered to have higher formant frequencies and a higher fundamental frequency, i.e. to resemble the productions of children. Listeners in those experiments were more likely to label intermediate /θ/-like tokens as /s/ when appended to a ‘child-like’ vowel than when it was appended to an ‘adult-like’ vowel—exactly the opposite of the pattern shown in Figure 5. That is, the listeners in those studies seemed more willing to interpret a /θ/-like token as an acceptable production of /s/ when they thought it was a child. Munson (2009) showed that this tendency was exaggerated when the listeners were told that they were listening to talkers who varied in age relative to a group that was told they were listening to adult talkers who varied in their height.

Discussion

The results of this experiment showed that people's perception of the accuracy of /s/ could be affected by experimental manipulations designed to induce different talker percepts. Moreover, the direction of this effect is much more complex than the simple effect of biasing intermediate productions that we hypothesized. At least some of the patterns noted here are likely due to the different types of information (including information about developmental variation in children's productions and sociolinguistic variation in adults' productions) that adults associated with frontal variants of /s/, as discussed in the previous section.

Finally, one might wonder whether experience mediates (and, ideally, attenuates) the effects of bias on ratings. The participants in the experiment in this section were diverse with respect to their experience hearing children's speech. Indeed, this diversity is by design, as these conditions were conducted as part of a larger computational-modelling project in which we intend to use these ratings as measures of the kind of feedback that children would receive during acquisition. Those of us who have either taken phonetics classes or who have both taken and subsequently taught phonetics classes know that the process of learning phonetic transcription is a long one. It typically involves many weeks of drill and practice in which students must simultaneously ignore the merely quasi-phonemic spelling system of English, and explicitly attend to fine acoustic detail that they previously processed only tacitly.

One would hope that the result of this extensive training would be reduced bias. Two pieces of evidence suggest that this is not the result. First, Schellinger et al.'s (2008) experiment compared the performance of less-experienced listeners (university undergraduates) to more-experienced ones (students in a graduate program in speech-language pathology). The two group's performance was statistically equivalent. Second, as summarized by Kent (1996), experience doesn't always mean reduced bias. Indeed, it often leads to *increased* bias, due presumably to the existence of a richer and more-entrenched set of expectations about how people ought to speak.

How I Learned to Stop Worrying and still love Phonetic Transcription

Imagine now the alien anthropologists years after they started their study of life on Earth. The linguistic anthropologists would have likely developed protocols for studying speech that involve detailed instrumental studies of articulation and acoustics, including perhaps extensive databases of productions collected with a consistent protocol. Given the unlimited resources that these aliens seem to be endowed with, we imagine that a separate research group would have spent an equivalent amount of time studying one other facet of human behaviour, our work-lives. These alien sociologists would likely have noted that humans who work with spoken language on a daily basis—speech-language pathologists, first- and second-language teachers, reading specialists, and audiologists, among others—typically work in settings where resources are much more limited. These poor Earthlings simply don't have the time or money or equipment to conduct the kind of detailed instrumental analyses of speech that the alien investigators do. The alien anthropologists and alien sociologists would have arrived at essentially the point where we humans are now: there is a disconnect between what we know about the sound structure of language, and how we can use that knowledge in our practice.

We imagine that readers of this article might be a bit dismayed by how sharp the divide is. Who can blame them? We have thus far painted a somewhat pessimistic picture of phonetic transcription. What's a clinician or a field researcher to do? Are we suggesting that we all need to give up phonetic transcription and rely solely on acoustic analysis and perception experiments? How are we going to describe the consonant inventories of typically developing children and children with speech sound disorders without phonetic transcription? How can we even do something as simple as providing a child with feedback on whether his or her

production is correct or incorrect in a therapy session without phonetic transcription? Have no fear. We are not suggesting that we must give up phonetic transcription. Rather, the point of this article is to remind researchers and clinicians again of some of the problems inherent to phonetic transcription. In addition, we'd also like to propose a simple modification to the usual transcription procedure and the adoption of some additional methods of evaluating children's speech.

One solution to this problem was developed for the transcription manual for the *παιδολογος* cross-linguistic database of phonological acquisition. The transcribers were given an additional option beyond the usual options of correct production, substitutions, and distortions. They were also trained to transcribe intermediate categories – productions that were intermediate between two sounds – using ordered combinations of the IPA symbols. The existence of intermediate productions has been noted even outside of the literature on covert contrast in the acoustic representations of sounds. For example, Pye et al (1988) noted that these are the productions that are the locus of most inter-transcriber disagreements and Stoel-Gammon (2001) suggested that transcribers label them as ‘fuzzy’. It turns out that both our trained phoneticians and our naïve listeners were remarkably good at identifying intermediate productions, as can be shown in Figs. 1 and 2 above. The naïve listeners differentiated between clear substitutions and intermediate productions for both the /d/-/g/ and the /s/-/θ/ experiments. In fact, listeners rated the intermediate [d]:[g] stimuli as less /d/-like than the clear-cut [d] for /g/ substitutions and as more /d/-like than the intermediate [g]:[d] substitutions. Similar results were found for the intermediate productions in the /s/-/θ/ experiment. These results suggest that ‘intermediate’ is a reliable transcription category.

Moreover, we encourage clinicians and field researchers to use the kinds of continuous rating scales that we have used in our research, such as those described in Urberg-Carlson et al. (2008). As Urberg-Carlson and colleagues described, these rating scales, particular Visual Analog Scales, are well correlated with acoustic parameters. These rating scale judgments can easily be implemented in both field research on phonological acquisition and in the clinic. More generally, we encourage spoken-language practitioners to see phonetic transcription as what it clearly is: an invaluable tool to help interpret the continuous physical speech signal. We further encourage clinicians who use the IPA to consider more closely the context in which the IPA was developed and in which it has changed. As discussed in depth by Ladd (in press), the IPA was designed in the late 19th century, long before the variation in acoustic phonetic detail presented in this paper had been studied, or even could have been studied. The IPA was simply not developed with the type of insights discussed in this article in mind. It behoves practicing clinicians and researchers to change their practices as the state of knowledge has changed.

Our work is by no means the only transcription system that endeavours to break the mold of how transcription is conventionally done. For example, the Multilayered Transcription system, described in Müller (2006), highlights the need to consider segmental production concurrent with other behaviours relevant to speech. Though the specific ways in which our system and Müller's system propose to overcome the limitations of conventional transcription are different, both are illustrations of the fact that clinicians and field researchers need not be bound by the practices that we were trained with.

Conclusion: Honouring our Colleague's Memory

We end this commentary by once again invoking its inspiration, Adele Miccio, and the conversation that led us to pick this topic. The point that Adele emphasized in this conversation was that transcription systems should not be composed of arbitrary symbols that serve different needs. If laterally misarticulated /s/ sounds produced by English-acquiring children are

identical to productions of the voiceless lateral fricative of Welsh, then the same symbol should be used to transcribe them. Phonetic symbols, she argued, shouldn't be used to reify a distinction that doesn't exist. They should be a tool—one of many—that we use to analyse speech. As such, they should serve the goal of helping us understand speech, including understanding typological diversity in speech, documenting developmental universals, or investigating some other topic, the same extensive goals of our fictional alien anthropologists.

Elsewhere in this issue are articles remembering Adele Miccio by writing on the specific topics that she worked on, particularly her seminal work on the relationship between stimulability and phonological development and disorders. We have chosen to honour her through a topic less directly related to her work, because it is a topic that we know she cared about deeply. Moreover, we know that she would continue to approach this topic with an open mind. We can imagine, for example, that some day we might find just how the laterally misarticulated /s/ of English is qualitatively different from the voiceless lateral fricative of Welsh, and see that the representations of those sounds should be faithful to that difference. We imagine that Adele Miccio would heartily embrace such a system, as doing so would be consistent with her life's goal of furthering our understanding of spoken language.

Acknowledgments

This research was supported by NSF grant BCS0729277 to Benjamin Munson, University of Minnesota Undergraduate Research Partnership Program grant to Marie K. Meyer and Benjamin Munson, and NIH grant R01 DC02932 and NSF grant BCS0729140 to Jan Edwards. We generously thank Kari Urberg-Carlson and Eden Kaiser for help with subject testing, and Jeff Holliday and Fangfang Li for help with the acoustic analyses in Table II.

References

- Arbisi-Kelm, T.; Beckman, ME.; Kong, E.; Edwards, J. Psychoacoustic measures of stop production in Cantonese, Greek, English, Japanese, and Korean. Paper presented at the 156th Meeting of the Acoustical Society of America; Miami. 10-14 November 2008; 2008.
- Ball, MJ.; Müller, N. *Phonetics for Communication Disorders*. Mahwah, NJ: Erlbaum; 2005.
- Baum SR, McNutt JC. An acoustic analysis of frontal misarticulation on /s/ in children. *Journal of Phonetics* 1990;18:51–63.
- Boersma, P.; Weenink, D. Praat v 5.0.35 [Computer Program]. Amsterdam: 2009.
- Dart, S. *Articulatory and Acoustic Properties of Apical and Laminal Articulations*. PhD Dissertation; UCLA Working Papers in Phonetics; Los Angeles: Department of Linguistics, University of California; 1991. p. 1-155. Reprinted as
- Drager. Speaker age and vowel perception. *Language and Speech*. in press. In press.
- Edwards J, Beckman ME. Methodological questions in studying consonant acquisition. *Clinical Linguistics and Phonetics* 2008;22:937–956. [PubMed: 19031192]
- Forrest K, Weismer G, Hodge M, Dinnsen DA, Elbert M. Statistical analysis of word-initial /k/ and /t/ produced by normal and phonologically disordered children. *Clinical Linguistics and Phonetics* 1990;4:327–340.
- Haggard M, Summerfield Q, Roberts M. Psychoacoustical and cultural determinants of phoneme boundaries: evidence from trading f0 cues in the voiced-voiceless distinction. *Journal of Phonetics* 1981;9:49–62.
- Hewlett N, Waters D. Gradient change in the acquisition of phonology. *Clinical Linguistics and Phonetics* 2004;18:523–533. [PubMed: 15573488]
- Kent R. Hearing and Believing: Some Limits to the Auditory-Perceptual Assessment of Speech and Voice Disorders. *American Journal of Speech-Language Pathology* 1996;5:7–23.
- Kerswill P, Wright S. The validity of phonetic transcription: Limitations of a sociolinguistic research tool. *Language Variation and Change* 1990;2:255–275.

- Kong, E. Unpublished PhD Dissertation. Columbus, OH: Department of Linguistics, Ohio State University; 2009. The Development of Phonation-type Contrasts in Plosives: Cross-Linguistic Perspectives.
- Langstrof C. Acoustic evidence for a push-chain shift in the Intermediate Period of New Zealand English. *Language Variation and Change* 2006;18:141–164.
- Ladd, DR. Phonetics in phonology. In: Goldsmith, J.; Riggle, J.; Yu, A., editors. *Handbook of Phonological Theory*. Malden, MA: Blackwell; in press To appear
- 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]
- 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]
- Maxwell EM, Weismer G. The contribution of phonological, acoustic, and perceptual techniques to the characterization of a misarticulating child's voice contrast for stops. *Applied Psycholinguistics* 1982;3:29–43.
- Müller, N., editor. *Multilayered Transcription*. San Diego: Plural Publishing; 2006.
- Munson, B. On Voiceless Fricative Perception: Vocal-Tract Normalization, and Sociointerindexicality. Oral presentation at the International Phonetics and Phonology Forum; Kobe University, Japan. August 26, 2009; 2009.
- Munson, B.; Zimmerman, L. Perceptual Bias and the Myth of the 'Gay Lisp'. Paper presented at the 2006 ASHA Convention; Miami. 16 November, 2006; 2006.
- Pye C, Wilcox KA, Siren KA. Refining transcriptions: the significance of transcriber 'errors.' *Journal of Child Language* 1988;15:17–37. [PubMed: 3280577]
- Repp B. Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants. *Language and Speech* 1979;22:173–189. [PubMed: 513908]
- Schellinger, S.; Edwards, J.; Munson, B.; Beckman, ME. Assessment of phonetic skills in children 1: Transcription categories and listener expectations. Poster presented at the 2008 ASHA Convention; Chicago. 20–22 November 2008; 2008.
- Stoel-Gammon C. Transcribing the Speech of Young Children. *Topics in Language Disorders* 2001;21:12–21.
- Tyler AA, Figurski GR, Langdale T. Relationships between acoustically determined knowledge of stop place and voicing contrasts and phonological treatment progress. *Journal of Speech and Hearing Research* 1993;36:746–759. [PubMed: 8377487]
- Urberg-Carlson, K.; Kaiser, E.; Munson, B. Assessment of children's speech production 2: Testing gradient measures of children's productions. Poster presented at the 2008 ASHA Convention; Chicago. 2008. p. 20–22.

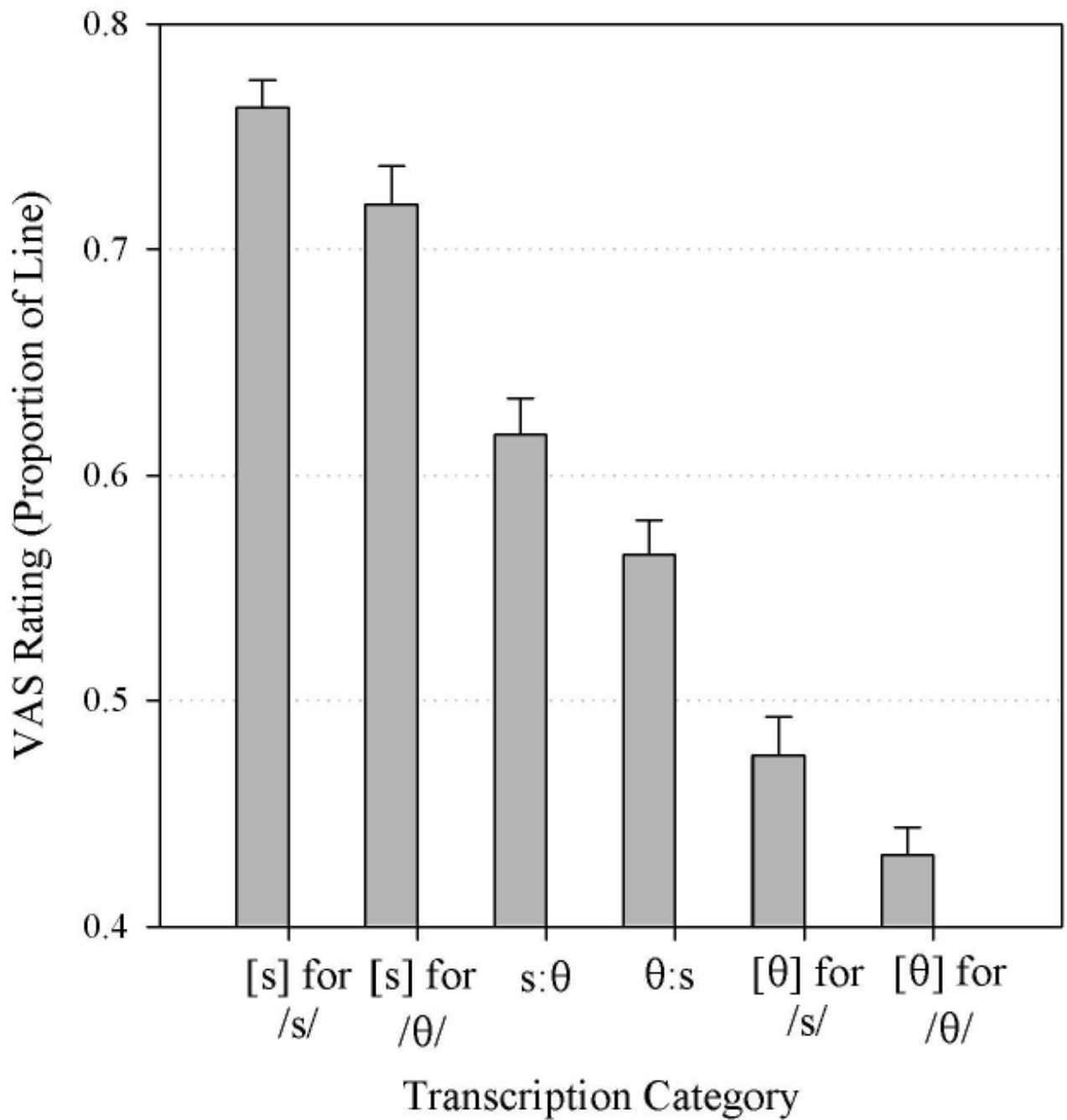


Figure 1. VAS ratings plotted against transcription category for the contrast between /s/ and /θ/. Dashed line represents the mid-point of the VAS scale.

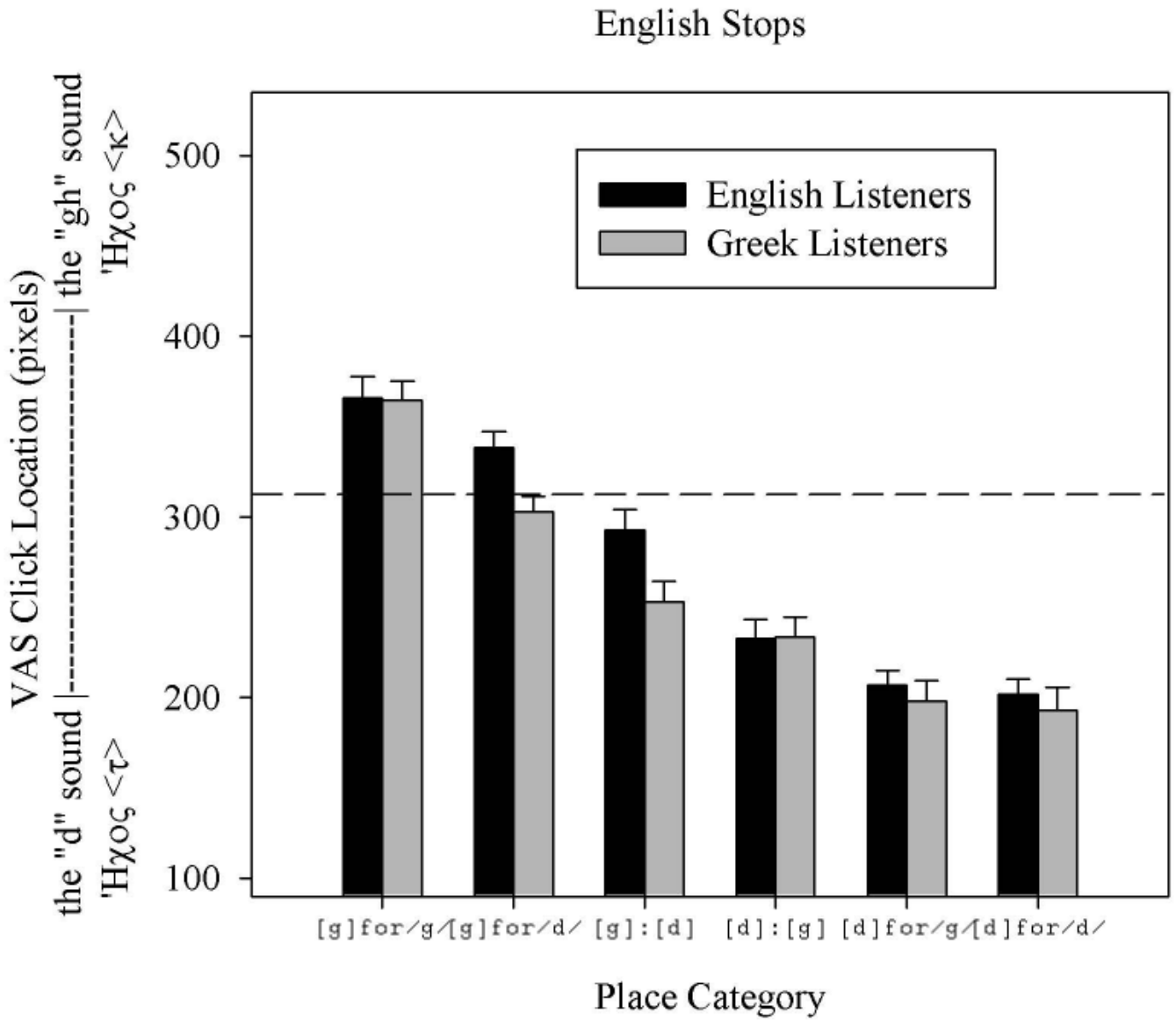


Figure 2. VAS ratings plotted against transcription category for the contrast between /d/ and /g/. Greek-speaking and English-speaking listeners are plotted separately. Dashed line represents the mid-point of the VAS scale.

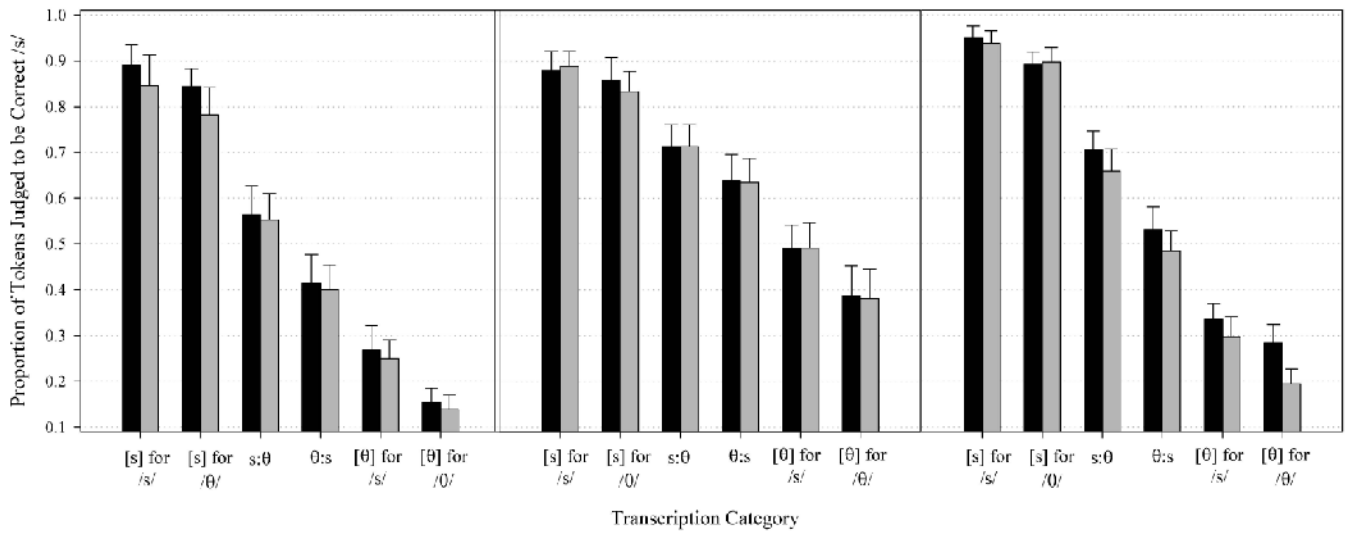


Figure 3. Proportion of ‘yes’ responses to the question ‘Is this an /s/’ in tasks 1 (left), 2 (middle), and 3 (right), plotted separately for the ‘really like’ carrier phrase (black bars) and the ‘weawwy yike’ carrier phrase (grey bars).

Table I

Summary of experimental conditions

Experimental conditions	Carrier phrases blocked by condition	Instructions mentioned ‘developmental misarticulations’	Carrier phrases matched CV sequences in f0
Schellinger et al.	no	yes	no
Condition 1	yes	no	no
Condition 2	yes	yes	no
Condition 3	yes	yes	yes

Table II

Acoustic characteristics of the stimuli.

Measure	[s] for /s/		[s] for /θ/		s:0		0:s		[θ] for /s/		[θ] for /θ/	
	Avg.	SD	Avg.	SD	Avg.	SD	Avg.	SD	Avg.	SD	Avg.	SD
<i>N</i>	50		24		26		30		24		46	
Peak ERB ^a	34.6	1.1	34.2	1.6	34.4	1.5	32.9	1.4	26.9	1.6	25.5	1.1
Compactness Index ^a	0.32	0.01	0.30	0.01	0.23	0.01	0.23	0.01	0.20	0.01	0.20	0.01
Total Loudness (sones) ^a	0.81	0.04	0.86	0.05	0.82	0.05	0.83	0.05	0.69	0.05	0.55	0.04
Duration (ms) ^b	209	9	210	13	214	12	223	11	187	13	174	9
Vowel F2 at onset (ERB)	21.9	0.2	22.1	0.3	22.0	0.3	21.7	0.3	21.6	0.3	22.0	0.2
Vowel f0 at midpoint (ERB)	7.5	0.1	7.3	0.2	7.5	0.2	7.6	0.2	7.3	0.2	7.4	0.2

^a $F[5,194] > 7.8, p < 0.001$.^b $F[5,194] = 3.3, p = 0.007$.