Effects of the distribution of acoustic cues on infants’ perception of sibilants

Alejandrina Cristià, a,n Grant L. McGuire b, Amanda Seidl c, Alexander L. Francis c

a Laboratoire de Sciences Cognitives et Psycholinguistique, EHESS, ENS-DEC, CNRS, Paris 75005, France
b University of California at Santa Cruz, Santa Cruz, CA 95064, USA
c Purdue University, West Lafayette, IN 47901, USA

A current theoretical view proposes that infants converge on the speech categories of their native language by attending to frequency distributions that occur in the acoustic input. To date, the only empirical support for this statistical learning hypothesis comes from studies where a single, salient dimension was manipulated. Additional evidence is sought here, by introducing a less salient pair of categories supported by multiple cues. We exposed English-learning infants to a multi-cue bidimensional grid ranging between retroflex and alveolopalatal sibilants in prevocalic position. This contrast is substantially more difficult according to previous cross-linguistic and perceptual research, and its perception is driven by cues in both the consonantal and the following vowel portions. Infants heard one of two distributions (flat, or with two peaks), and were tested with sounds varying along only one dimension. Infants’ responses differed depending on the familiarization distribution, and their performance was equally good for the vocalic and the frication dimension, lending some support to the statistical hypothesis even in this harder learning situation. However, learning was restricted to the retroflex category, and a control experiment showed that lack of learning for the alveolopalatal category was not due to the presence of a competing category. Thus, these results contribute fundamental evidence on the extent and limitations of the statistical hypothesis as an explanation for infants’ perceptual tuning.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

Research on the development of speech suggests that, in early infancy, humans show speech discrimination abilities that do not appear to depend on their language experience; however, by the end of their first year, their perception is more tuned to the sounds present in the ambient language (Jusczyk, 1997). In this paper, we present evidence providing moderate support to the hypothesis that, if this tuning is due to category learning, then it may be explained as a result of attention to statistical distributions of acoustic cues in the speech infants hear.

More specifically, we first review, in Section 1.1, previous findings on infants’ perceptual acquisition, which suggest that this process involves the formation of categories on the basis of pre-existing auditory-perceptual abilities (as proposed by e.g., Aslin & Pisoni, 1980; Kuhl et al., 2008), and not simply the selection of a subset of categories among an innately given set (as suggested by Gervain & Werker, 2008; Liberman & Mattingly, 1985 among others). If perceptual acquisition involves learning, it is plausible that this process is aided by infants’ attention to the statistical
distributions of acoustic cues. Extant empirical evidence supporting this statistical learning hypothesis is summarized in Section 1.2. Nonetheless, the contrasts used in previous research relied on a single, psychoacoustically salient dimension. Therefore, there is still little evidence that the statistical learning hypothesis can scale up to the challenges infants face in the task of natural language acquisition, as argued in Section 1.3. This evidence was sought in two experiments, reported in Sections 2 and 3. These experiments show that infants’ perception is affected by acoustic cue distributions, but possibly only in regions in acoustic space to which infants are already sensitive. The implications of these and previous findings are discussed in Section 4.

1.1. Infants’ perception: selection versus bounded statistical learning

It is commonly reported that infants are able to discriminate contrasts that do not exist in their ambient language. For example, Japanese 6-month-old infants can discriminate the non-native contrast [r-I] (Kuhl et al., 2006), a contrast that is remarkably difficult for their elders to hear (Iverson et al., 2003). A second fact of infant perception is that, by about 12 months of age, monolingual infants’ sensitivity tends to be maintained or improved for contrasts present in their ambient language, and to decline for others that are not functional in that language.
English- and French-learning 6–8-month-olds perform universal discrimination abilities are not perfect. For example, ambient language. are potentially linguistically relevant, and acquisition consists of the ability to discriminate any, all, and only phonetic changes that among an innately given feature set (e.g., Gervain & Werker, 2008; Liberman & Mattingly, 1985). That is, infants are born with the ability to discriminate any, all, and only phonetic changes that are potentially linguistically relevant, and acquisition consists of losing the discrimination abilities that are not required by the ambient language.

However, it is important to remember that early (so-called universal) discrimination abilities are not perfect. For example, both English- and French-learning 6–8-month-olds perform poorly with [d–θ] discrimination (Polka, Colantonio, & Sundara, 2001), and neither Filipino nor English-learning 6–8-month-olds can discriminate [n–ŋ] (Narayan, Werker, & Speeter Bedder, 2009). Moreover, changes in infants’ discrimination skills at the end of the first year are also not a matter of all-or-none. First, patterns of decline appear to be modulated by the frequency of the sounds; for instance, Anderson, Morgan, and White (2003) argue that English-learning infants lose the ability to discriminate the Hindi [t̪-t̪] contrast earlier than they do the Ñthlakampx [k̪-'q̪] contrast because coronals are more frequent in English. Second, in some cases, infants’ sensitivity for contrasts present in the target language remains relatively poor by the end of the first year. This is the case for [d–θ] (Polka et al., 2001), and [s–ʃ] (Nittouer, 2001) in English. Finally, decline is also modulated by the articulatory/perceptual characteristics of the sounds, such that some sounds are still discriminable in late infancy in spite of the absence of experience with them (Best & McRoberts, 2003; Best, McRoberts, LaFleur, & Silver-Isernstadt, 1995; Best, McRoberts, & Sithole, 1988).

These four sets of findings are hard to reconcile within the early view of acquisition as feature selection. Instead, Aslin and Pisoni (1980) offer a typology of four possible developmental changes in infant perception that can be used to better understand the findings reported above. When discrimination abilities are robust in the absence of experience, subsequent exposure to a language having that contrast would result in (1) maintenance (as in the English [b–v] contrast: The sensitivity of both French and English learners to this contrast at 6–8 is indistinguishable from that of 10–12-month-olds; Polka et al., 2001), while an initially robust sensitivity that declines if the ambient language does not contain the relevant contrast would be a case of (2) attenuation (as in American infants’ sensitivity to the Hindi [t̪-t̪] contrast; Anderson et al., 2003). In contrast, when discrimination abilities are initially weak, exposure to a language that recruits that contrast should result in (3) enhancement (perhaps exemplified in American children’s discrimination of English [r–l]; Kuhl et al., 2006), or possibly (4) induction, in order to achieve native perception (as in Filipino infants’ perception of the [n–ŋ] contrast: At 6–8 months infants fail to discriminate this contrast, but succeed at 10–12 months; Narayan et al., 2009). To this typology, we add the two cases of inelasticity mentioned above: (5) poor initial sensitivity that is not improved by the presence of the sound (the [d–θ] contrast is present in English and absent in French, yet English- and French-learning infants show similar discrimination scores at both 6–8 and 10–12 months; Polka et al., 2001); (6) good initial sensitivity that is maintained despite the lack of evidence for the sound or contrast in the native language.

The latter includes clicks (Best et al., 1988) and certain vowels. (Polka & Bohn, 1996 found no differences in linguistic backgrounds, cf. Polka & Werker, 1994, who documented that 4- but not 6-month-olds dishabituated to the non-native [u–y] in English learners, but Cardillo, 2010, documented an improvement of performance with age with the same contrast.)

Given that these paths of development take place in the first year of life, before infants acquire a sizable lexicon (Caselli et al., 1995), it is unlikely to be lexical knowledge that drives them. Instead, recent research in other domains of language acquisition highlights infants’ abilities to learn from statistical patterns (see, for example, Aslin & Newport, 2009; Saffran, 2009, for recent reviews). Pierrehumbert (2003) has developed a version of this statistical learning hypothesis for phonetic acquisition, according to which infants may postulate (rudimentary) categories by keeping track of differential frequency distributions on the basis of their perception of the spoken input. That is, infants map their acoustic input onto a non-linear auditory space, which is warped by peaks and valleys of sensitivity. As they gain experience, clusters of exemplars shift this warped space into the perceptual map that best fits the target language’s distributions (as they are when converted into the non-linear auditory space). Notice that this statistical learning is bounded by perceptibility, because it does not assume that sheer input frequency will be immediately reflected in final perceptual abilities. Thus defined, bounded statistical learning could account not only for the 4 types of development predicted by Aslin and Pisoni (1980) on the basis of exposure to speech, but also for the two inelastic cases, under the assumption that these peaks in frequency happen to fall in auditory “blind spots” and the valleys of frequency in auditorily over-salient regions. Although this hypothesis provides a good fit with extant data from variation across contrasts in natural language acquisition, too many factors may be at play in those processes to be certain that a combination of input statistics and auditory sensitivities accounts for the typology of events found. Further evidence has been sought through controlled, laboratory-based learning studies, summarized in the next section.

1.2. Laboratory-training studies: limited evidence for the statistical hypothesis

Two laboratory-training studies document changes in perception as a result of exposure to different frequency distributions in acoustic space (Maye, Weiss, & Aslin, 2008; Maye, Werker, & Gerken, 2002). In those studies, Maye and colleagues investigated the effect of short exposure to different distributions of voice onset time (VOT) on infants’ perception of voiceless and aspirated stops (Maye et al., 2002) and prevoiced and voiceless stops (Maye et al., 2008). Specifically, 6- and 8-month-olds heard an 8-step continuum between the endpoints [t̪a-t̪a] in Maye et al., 2002; [da-ta] in Maye et al., 2008), whereby each step varied in frequency of occurrence. For infants hearing a bimodal distribution, steps 2 and 7, adjacent to the two endpoints, were presented frequently, and the center tokens 4 and 5 very infrequently. Contrastingly, for infants in the unimodal condition, steps 4 and 5 in the center were highly frequent, while the endpoint steps 2 and 7 were very infrequent. Crucially, infants in both conditions heard steps 3 and 6 the same number of times. Infants were then tested on their discrimination of these steps in different ways in the two papers: In Maye et al. (2002) discrimination was tested by comparing looking times to trials where the 2 steps alternated versus trials consisting of repetitions of one of the steps (Maye et al., 2002); in Maye et al. (2008) by assessing dishabituating to token 3 (closer to [d]) after habituation to token 6 (closer to [t]). Results showed that infants in the bimodal distribution

Author's personal copy

conditions discriminated steps 3 and 6 (looking times to alternating and non-alternating trials differed significantly; dishabitation generalized across step exposure), whereas infants in the unimodal condition did not (looking times to alternating and non-alternating trials did not differ; dishabitation did not generalize across step exposure).

While these results support the theory of statistical learning, infants’ learning situations in these 2 studies do not exhaust the challenges encountered in natural language acquisition. There are at least 2 factors that have not been covered, and yet they likely have an impact on learnability. One relates to the fact that contrasts themselves vary in salience, with those along the VOT continuum arguably being relatively salient, possibly due to auditory nonlinearities (Holt, Lotto, & Diehl, 2004), and it is not clear that evidence based on such robust contrasts can be extrapolated to less robust ones. Second, although many other acoustic parameters correlate with stop voicing in syllable-initial position (Lisker, 1986), adult listeners tend to depend primarily on VOT (Francis, Kaganovich, & Driscoll-Huber, 2008). However, it is possible that other sound contrasts may be more equally cued by multiple acoustic correlates, even for adults. In these cases, infants need to at least keep track of multiple acoustic correlates in order to achieve native perception. These arguments motivate the current research, and are therefore expanded upon in the following 2 subsections.

1.2.1. Salience

While the notion of “salience” appears to be intuitive, it is complicated to zero-in on the reasons why certain sounds, contrasts, and dimensions are more or less so (see e.g., Holt & Lotto, 2006, for a discussion). For the purposes of the present research on category learning, we adopt the following working definitions of salience: A contrast is salient if its members are robustly discriminated with no prior experience and/or if the contrast can be learned with little training. Second, a contrast is psychoacoustically salient if these discrimination abilities are shown by non-human animals and/or by humans when tested with non-speech stimuli. Finally, a sound within a contrast will be considered non-salient if it is (a) less frequent cross-linguistically than its counterpart (as an indirect proxy, since many factors could affect cross-linguistic frequency); and (b) it is often misidentified by both listeners whose native language has the contrast, and others whose native language does not (Babel & McGuire, 2010; Narayan, 2008).

As mentioned above, Maye and colleagues have investigated learning of stop voicing categories, which depend primarily on VOT. Extant research strongly suggests that contrasts along VOT fit our definition of salient. To begin with, infants with little experience exhibit categorical perception of the voiceless-aspirated contrast (Eimas, Siqueland, Jusczyk, & Vigorito, 1971; Lasky, Syrdal-Lasky, & Klein, 1975; Streeter, 1976), and the prevoiced-voiceless contrast (Aslin, Pisoni, Hennessy, & Perey, 1981; but see Lasky et al., 1975, for arguments that this contrast may be harder for infants than the voiceless-aspirated one). As for learning effects, improved performance has been documented for contrasts involving this dimension (e.g., within-category discrimination training, Pisoni & Lazarus, 1974). Furthermore, both speech and non-speech training studies with adults support the hypothesis that VOT discontinuities specifically facilitate category learning. For example, brief training suffices to re-learn American listeners on the prevoiced-voiceless contrast, which maps onto a single phonemic category in their ambient language (Pisoni, Aslin, Perey, & Hennessy, 1982; see also Tees & Werker, 1984, who document that voicing contrasts are easier to re-learn than place contrasts). Finally, laboratory learning of a category centered on VOT auditory discontinuities is remarkably difficult, much more so than learning of a category that does not span this region (Holt et al., 2004). Other research suggests that this salience is based on the psychoacoustic properties of VOT contrasts, as categorical perception has been demonstrated by both non-human animals (e.g., Kuhl & Miller, 1975, 1978; Kuhl & Padden, 1982), and by both adult and infant humans with non-speech continua (e.g., adults: Elangovan & Stuart, 2008; Pisoni, 1977; infants: Jusczyk, Pisoni, Walley, & Murray, 1980; Jusczyk, Rosner, Reed, & Kennedy, 1989). In view of this evidence, there is little doubt that, for some regions of acoustic space, there are discontinuities along VOT that are psychoacoustic in nature, and which contribute to linguistic phonetic contrasts that are salient.

Unfortunately for learners, few phonetic contrasts depend on VOT, whose relevance may be limited to signaling stop voicing. Yet infants must learn many other contrasts, including those for which salience has not been established. Given that salient contrasts are easier to learn, previous results may not extend to less salient categories. In order to be a tenable theory of the acquisition of speech perception over the first year, statistical learning needs to generalize to other documented cases, including those of initially weak sensitivities.

1.2.2. Multi-cue contrasts

It is clear that most speech sound contrasts are correlated with multiple aspects of the speech signal, many of which vary in synchrony. For example, there are 16 correlates (or acoustic characteristics) that tend to coincide with the voicing contrast of intervocalic stops in English (Kingston & Diehl, 1994; Lisker, 1986), each of which could potentially be useful in making voicing distinctions (and many of which are, in fact, integrated at some perceptual levels: Kingston, Diehl, Kirk, & Castleman, 2008). Nonetheless, VOT is evidently a salient cue, and it could be the main factor driving infants’ perception of syllable-initial voicing in the studies by Maye and colleagues, as it does for adults (e.g., Benki, 2005 carried out a logistic regression on voicing judgments, in which the β value for VOT was about 60 orders of magnitude larger than that of F1; this relative weight is increased under cognitive load, Gordon, Eberhardt, & Rueckl, 1993). However, other cases call for listeners’ attention to be more distributed across the multiple correlates, as in the [ca] (an alveopalatal sibilant followed by a low back vowel) and [sa] (a retroflex sibilant followed by a low back vowel) contrast, for which Mandarin talkers rely on acoustic correlates in the consonantal and the vocalic portions to a similar extent (McGuire, 2007; Chiu, 2010). To be able to learn certain speech categories, then, infants must track numerous acoustic correlates simultaneously in order to correctly identify and discern sounds, particularly in the case of multi-cue contrasts. Given that the voicing categories used in Maye et al. (2002, 2008) could have been resolved on the basis of a single correlate (VOT), those studies cannot address the question of whether infants are able to learn frequency distributions present along multiple correlates simultaneously. Multi-cue contrasts pose additional problems to the learner, according to a number of category training studies, primarily on human adults (speech: Goudbeek, Cutler, & Smits, 2008; non-speech: Goudbeek, Swingley, & Kluender, 2007; Goudbeek, Swingley, & Smits, 2009; but see Holt & Lotto, 2006; and visual: e.g., Alfonso-Reese, Ashby, & Brainard, 2002; Ashby, Queller, & Berretty, 1999; see Lea & Wills, 2008, for arguments that this “unidimensional bias” is also

---

1 We call such contrasts multi-cue, because listeners’ attention is distributed over multiple acoustic correlates; we reserve the use of the word ‘dimension’ for correlates or set of correlates that are manipulated together in a given study. Thus, a study where VOT is manipulated independently from one, F1 and F2 is both multi-cue and multidimensional; one where onset pitch and F1 are manipulated in tandem is multi-cue and unidimensional.
found in non-human animals). For example, Ashby et al. (1999) show that, in visual category learning, adults tend to rely on a single manipulated dimension unless otherwise prompted, and Goudbeek et al. (2009) report that, in the absence of feedback, adults quickly revert to unidimensional solutions after having learned multidimensional sound categories. Since early phonetic learning is necessarily unsupervised (that is, there is no corrective feedback to use multiple dimensions), this unidimensional bias is likely to be present in infancy. In other words, the question is whether, when faced with stimuli where multiple acoustic characteristics vary, infants can and do keep track of several of those correlates, a pre-requisite to noticing their co-occurrence. If infants, like adults and non-human animals, favor unidimensional/uni-correlate solutions, does this mean that they track only one correlate and ignore all others?

This question cannot be answered by previous research on laboratory-based learning or on natural language acquisition. Much research on subphonemic cues has focused on whether infants are able to discriminate when presented with a subset of cues, or whether the presence of multiple cues facilitates discrimination (e.g., burst spectrum and/or formant transitions as a cue to stop place; Miller, Morse, & Dorman, 1977; Moffitt, 1971; Walley, Pisoni, & Aslin, 1984; Williams & Bush, 1978; see also Jusczyk, 1981). Others have assessed whether infants compensate for subphonemic patterns (e.g., Eimas, 1985; Eimas & Miller, 1980a, 1991; Fowler, Best, & McRoberts, 1990; Levitt, Jusczyk, Murray, & Carden, 1988), in an attempt to see whether infants have innate access to gestural units (e.g., Fowler, 2006; Fowler et al., 1990), or whether such compensation responds to more basic auditory processes (e.g., Lotto, Kluever, & Holt, 1997; Lotto & Holt, 2006). Neither of these lines of research demonstrates developmental or experiential changes, so they do not bear on how infants learn multi-cue contrasts. A recent study provides the first piece of evidence to this effect: Ko, Soderstrom, and Morgan (2008) document that 14 but not 8-month-olds prefer to listen to trials where long vowels are followed by voiced stops and short vowels by voiceless stops, over trials with long vowels followed by voiceless stops and short ones by voiced stops. Thus, it appears that the learning of the co-occurrence of acoustic correlates (here, extrinsic vowel length and stop voicing) takes place around the end of the first year of life. Given the similarity in timeline with the perceptual tuning reviewed in Section 1.2, these results lend indirect support to the extension of the statistical learning hypothesis to multi-cue contrasts.

1.3. Current research

In short, since the statistical hypothesis is theoretically attractive as a parsimonious explanation to infants’ phonological acquisition, it is worthwhile to garner further empirical evidence concerning the likelihood that it may scale up to the challenges found in natural language acquisition. The present study was designed as one step in this direction, by extending previous work in two ways: using a non-salient contrast and varying multiple dimensions. The contrast between [ca] and [sa] as implemented in Polish fulfills both conditions (non-salience, and multi-cue), as explained below. It should be noted that the phonetic transcription of these Polish sounds has been subject to some debate (e.g., Ladefoged & Maddieson, 1996, described the alveolopalatal as palatalized post-alveolars), which now appears to have been resolved (see Nowak, 2006; Zygis & Hamann, 2003, for discussions).

1.3.1. Salience

The initial purpose of this study was to test infants’ category learning with sounds that were less salient than VOT, and other contrasts which very young infants (and sometimes non-human animals) can easily perceive (e.g. the place contrasts, [b-d], [b-g], [t]- and [k-q] in stops and [w]-glides, the manner contrasts [b-w], [b-m] and [r]-l: Bertoncini, Bijeljac-Babic, Blumstein, & Mehler, 1987; Eimas, 1975; Eimas & Miller, 1980a, 1980b; Jusczyk, 1977; Jusczyk & Thompson, 1978; Miller & Eimas, 1983; Morse, 1972; Werker & Tees, 1984; cf. Yoshida, Pons, Mage, & Werker, 2010). In contrast, evidence for infants’ discrimination of fricative place of articulation is mixed. For instance, Eilers, Wilson, and Moore (1977) report that both 6–8-month-olds and 12–14-month-olds fail to discriminate [t]-[l], while 2-month-olds (tested with a different method) succeeded with the same contrast, according to Levitt et al. (1988). Three-month-olds were unable to discriminate [sa]-[za], but they succeeded with [as]-[az] (Eilers, 1977; see also Aslin, Pisoni, & Jusczyk, 1983). While Holmberg, Morgan, and Kuhl (1977) are often cited as demonstrating that 6-month-olds can discriminate [s]-[f] (e.g., Jusczyk, 1997: p. 53; Levitt et al., 1988: p. 362), Nittroouer (2001) finds that few can. Using a within-subject design, Nittroouser (2001) tested infants between 6 and 14 months of age on the fricative place and either a vowel contrast or stop voicing contrast. She reports that only 6 out of 15 infants who could distinguish vowel quality (either [sa-su] or [za-fu]) could also discriminate [sa], while out of 8 infants who could distinguish a stop voicing contrast ([ta-da]), none discriminated the sibilants.

Although the second criterion for salience cannot be evaluated (no previous research has focused on non-human animals’ discrimination of fricatives), the above results suggest that sibilant place of articulation does not fit our first criterion for salience, since most evidence indicates that they are not easily discriminated in the absence of experience. Furthermore, while no previous work documents infants’ discrimination of [ca-sa], this contrast is likely more difficult than the [sa-ja] one. First, [c-s] are about 10 times less frequent than, for instance, [s-f] (in Maddieson, 1984: IDSISP, [s]: 21 languages; [c]: 9 languages; an additional 2 languages contain both: [s]: 197 languages; [f]: 189 languages; of which 82 languages have both). Furthermore, the contrast likely spans a smaller phonetic distance, since [s] and [f] are the extremes in sibilant place (Gordon, Barthmaier, & Sands, 2002). This intuition is supported by acoustic measurements from languages having [c-sa], and the acoustic distance between [s] and either [c] or [f] is larger than that between [c] and [s]; in short, [s] ≠ [c] ≠ [f] (e.g., Jassem, 1979; Kudela, 1968; Nowak, 2006; Zygis & Pargett, 2010). The reduced acoustic distance between [c] and [s] may be particularly problematic for Polish, as they are more acoustically similar in Polish than in the Mandarin case (e.g., Li, 2008; see Chiu, 2009, for a direct comparison between the two).

Given that [c] is less frequent cross-linguistically than [s], one may wonder whether it is actually less salient than the other fricatives. Indeed, [c] is often misidentified, according to data reported in Nowak (2006). When faced with a variety of cross-spliced stimuli, Polish speakers perform worse with [c] even when all cues are available (calculated by the present authors from Table 5 in the original paper: Mean difference 7.7, pooled SE=5.61; t(7)=−2.57; p<.05). This pattern of results has been replicated with Mandarin Chinese listeners (a language that also has dental, alveolopalatal and retroflex sibilants), who show lower sensitivity, higher response times, and higher biases for alveolopalatals than retroflexes both for Mandarin stimuli and for Polish stimuli (Chiu, 2010). Moreover, when confronted with a bidimensional grid between the Polish retroflex and alveolopalatal, Mandarin listeners tend to label a majority of the tokens as
retroflex (McGuire, 2007: p. 74). Finally, in the same grid identification task, American English listeners, who have no phonemic experience with [c-s], labeled tokens near the retroflex of the grid more consistently than those at the alveolarpalatal end (McGuire, 2007; this was evident in several experiments; see e.g., p. 51 and p. 73). In short, Polish [s] appears to be perceptually more salient than Polish [c] for Polish listeners as well as for non-native listeners, both for those who have phonemic experience with these sounds (Mandarin) or those who do not (English).

1.3.2. Multiple cues

Acoustic and perceptual studies suggest that the identity of Polish retroflex and alveolarpalatal sibilants in syllable-initial position depends on a number of acoustic correlates, which are distributed across the frication and vocalic portions. The two sounds are reported to differ on two major acoustic properties: the distribution of energy across the frequency spectrum during the period of fricative noise; and the pattern of formant transitions (in particular the second formant, F2) immediately following the onset of voicing (e.g., Lisker, 2001). Polish listeners’ identification is affected by cues in the vocalic portion (Nowak, 2006; Experiment 1). Similarly, Mandarin listeners attend to both the frication and the vocalic portions in their identification (Chiu, 2010) and discrimination (McGuire, 2007) of the Polish sibilants (the same as they do for their own sibilants, Chiu, 2010). Thus, manipulating the consonant and vowel dimension separately will lead to a bidimensional, multi-cue contrast.

1.3.3. Summary of the motivation and predictions

The statistical learning hypothesis predicts that infants’ perception will be affected by distributions of acoustic cues in their input, and that this may form the basis of category learning. Previous research has shown that infants’ perception is affected when the contrast to be learned relies on a single, salient acoustic dimension, VOT. The present study sought to put this hypothesis to a more stringent test by comparing infants’ perception of two non-salient sounds across different exposure conditions, which differed in the frequency distribution of acoustic correlates spread over two different portions of the signal (since consonantal and vocalic portions were manipulated separately). The stimuli chosen were based on the non-salient and multi-cue contrast found in Polish syllables [ca-sa]. However non-salient these sounds may be, following the statistical learning hypothesis, we predict that infants’ perception will be different following the different exposure conditions. In contrast, if frequency distributions do not influence infants’ perception in non-salient or multi-cue contrasts, no difference will be found across exposure conditions.

2. Experiment 1: Flat and Two Peak distributions

In the first experiment, two groups of infants heard one of two different distributions during an initial exposure period. In the Flat group, infants heard all familiarization tokens repeated the same number of times, such that no acoustic category would be promoted by the input. In the Two Peak group, infants heard the “natural corners” of the grid more frequently. After this initial exposure, all infants were tested using the Headturn Preference Procedure (Jusczyk & Aslin, 1995; used with 4-month-olds in e.g., Cristià, Seidl & Gerken, in press; Seidl & Cristià, 2008; Seidl et al., 2009) on their perception of tokens in different areas of the grid. This is unlike previous research, where one of two designs was used. During the test phase, infants in Maye et al.’s (2002) were presented with 2 types of trials: Alternating trials, in which two tokens, one for each category, were played in succession; and non-alternating trials, in which the same tokens were presented in 2 separate trials. The authors then collapsed across the 2 non-alternating trials and compared infants’ looking times of this average with that to the alternating trials, under the assumption that infants would respond to the variability within trials. However, if one of the categories is more salient than the other, then the non-alternating trials may not be comparable, and looking times should not be averaged across them. Indeed, Maye et al. (2008) did not repeat this procedure for learning of prevoiced-unaspirated, given that there could be differences in the perception of these two sounds. That is, in an experiment using the Conditioning Head-Turn procedure, Aslin et al. (1981) documented that the prevoiced served as a better “background” than the unaspirated one, a pattern that has since been associated with perceptual asymmetries (Polka & Bohn, 2003), such that perceptually stronger categories are worse backgrounds (similarly to what happens with the perceptual magnet effect, e.g., Kuhl et al., 1992, and discussion in Polka & Bohn, 2003: p. 222 and p. 227). Therefore, Maye et al. (2008) used a habituation–dishabituation paradigm, in which the stronger category (the unaspirated stop) served as background or habituation stimulus, and a looking time contrast was expected for the dishabituation stimulus.

There are two important problems with this second procedure that prevented us from adopting it. First, throughout habituation infants hear additional exemplars of one of the categories, thus modifying the distribution that they are exposed to. Even though this manipulation did not affect infants’ perception in Maye et al. (2008), one cannot be certain that it would have an equally null effect with categories varying in acoustic distance and salience, such as the sibilants used here. The second problem with the habituation–dishabituation test is that it assumes that discrimination of the two tokens belonging to the different categories is a sign of learning of both categories. However, a similar result could ensue if only one category is learned, as long as the other token being presented is different enough to be seen as a bad exemplar or an outlier of the learned category. For example, Kuhl et al. (1992) show that infants are less able to make discriminations close to the prototype, but substantially better when discrimination involves less prototypical tokens. Consequently, there is an alternative interpretation of Maye et al. (2008), according to which infants learned only one category, that of unaspirated stops. Clearly, this interpretation does not invalidate Maye and colleagues’ conclusion that infants’ perception was altered by the distribution of acoustic cues they heard; on the contrary, it is clear that their perception was affected by acoustic cue distributions. However, we do believe that the test adopted cannot document learning of both categories. Therefore, we opted for a design that allowed us to assess learning within each category independently, by measuring looking times to combinations of correlates that are more or less frequent in the infants’ initial exposure. This was encoded in the variable Place for statistical analyses.

Another important choice of the design concerned the stimuli. A unidimensional continuum along which two or more correlates covaried would indeed be multi-cue. However, in such a design infants would have been free to still resolve the task on the basis of a single correlate. Specifically, a great deal of work shows that children and adults do not attend to all correlates equally, but pay more attention to some and less to others (Francis, Baldwin, & Nusbaum, 2000; Francis et al., 2008; Gouldbeek & Swingley, 2006; Holt & Lotto, 2006; Mayo, Scobbie, Hewlett, & Waters, 2003; Mayo & Turk, 2004, 2005; Nittrouer, 1992, 2002, 2006; Nittrouer, Miller, Crowther, & Manhart, 2000; Ohde, Haley, & McMahon, 1996; Sussman, 2001; Wagner, Ernestus, & Cutler, 2006). Similarly, infants could have paid attention to only one salient acoustic correlate, and learned only its distribution, in which case the contribution of the present paper would be limited to an extension to a non-salient contrast. Both stimuli and testing methods were designed to avoid this confound. In particular, consonant and vowel portions were manipulated independently; infants were exposed to stimuli where the consonantal and
vocalic dimensions of the stimuli were correlated during initial
exposure, but decorrelated during test. In particular, during test,
stimuli varied on only one of the two dimensions; this way, infants
had to, at least, have kept track of one correlate within each of the
two dimensions during the previous exposure to succeed during test.
That is, let us imagine that infants pay attention only to the formant
transitions encoded in the vocalic portion. In this case, infants will
learn the distributions of the vocalic portion only, and will succeed
with the stimuli varying along the vocalic dimension. However, they
will show no effect of exposure distribution when tested with the
consonantal dimension, since they have not attended to the cues in
the consonantal portion. In contrast, if they pay attention to at least
one correlate in each manipulated dimension, and additionally they
can make perceptual distinctions when one dimension is held
constant, then they will show effects of exposure in both dimensions;
or at least, there will be no interaction Distribution Dimension.
Incidentally, notice that no work documents an overwhelming bias
at this age towards static/dynamic or consonantal/vocalic portions
(see e.g., Bohn & Polka, 2001).

2.1. Methods

2.1.1. Participants

Sixty-four (32 in each condition) English monolingual, fullterm
infants were included (M=5.0 months, range 3.95–6.02 months,
30 female). An additional 34 infants were not included for the
following reasons: failing to finish the experiment due to fussing,
crying or falling asleep (16); experimenter or equipment error (6);
being exposed to a language other than English (3); being
premature (1); or having looking times shorter than 1 s on any
given trial (8).

As noted above, using a non-salient contrast and varying multi-
ple dimensions may make learning more difficult. In order to
maximize the chances of success in this unfavorable scenario, we
tested 4–6-month-old infants, given that mounting evidence sug-
gests that infants’ phonetic learning abilities become increasingly
constrained with experience. For example, younger infants succeed
at learning sound patterns that older infants do not detect (Cristià &
Seidl, 2008; Cristià, Seidl, & Francis, in press; Cristià, Seidl, & Gerken,
in press; Gerken & Bollt, 2008; Seidl et al., 2009); and 10-month-
olds fail to learn a VOT contrast that 8-month-olds acquire (Yoshida
et al., 2010). Infants tested here were therefore younger than those
in the previous infant category learning studies (Maye et al., 2002:
6- and 8-month-olds; Maye et al., 2008: 7–9-month-olds; Yoshida
et al., 2010: 10-month-olds).

2.1.2. Stimuli

The stimuli presented both in the initial exposure and testing
were produced by modifying a pair of syllables [ca] and [sa]
produced by a male Polish speaker. The original syllables were
recorded in a sound-shielded booth with a head-mounted micro-
phone (AKG, model C420) and a Marantz PMD670 solid state
recorder at 44.1 kHz sampling rate and stored in .wav format.
These original syllables were selected on the basis of clarity as
well as similarity to the acoustic characteristics for Polish alveo-
lopalatal and retroflex sibilants reported in Nowak (2006). The
acoustic measurements for the original syllables are reported in
Table 1 and a graphic depiction of the most important acoustic
parameters in the fricative and vocalic portions are presented in
Figs. 1 and 2.

Even though the main correlates of sibilant place identity are
the centroid of the distribution of energy during the frication and

<table>
<thead>
<tr>
<th></th>
<th>[sa]</th>
<th>[ca]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fricative spectrum peak (Hz)</td>
<td>2890</td>
<td>3890</td>
</tr>
<tr>
<td>Onset F2 (Hz)</td>
<td>1420</td>
<td>1720</td>
</tr>
<tr>
<td>Midpoint F2 (Hz)</td>
<td>1240</td>
<td>1320</td>
</tr>
</tbody>
</table>

Fig. 1. Graphic representation of the most relevant acoustic cue in the fricative continuum, the noise spectrum, as rendered by cepstral smoothing with a 500 Hz bandwidth. The darkest line represents the alveolopalatal end, the dotted line the retroflex end, and the light gray lines the intermediate tokens f3 and f6.
onset F2 in the vowel, it is clear that other correlates may have a perceptual effect. For example, Nowak (2006) finds that Polish listeners' identification is affected not only by formant transitions early in a following vowel, but also, to some extent, by more distant cues in the vowel. Since synthesizing simplified stimuli (e.g., a pole followed by synthetic F1–F3) would necessarily rely on assumptions regarding which cues are perceptually relevant to infants, we chose an alternative method of stimuli generation. We split the [ca] and [sa] syllables into a frication portion and a vocalic portion, and generated one continuum for each one of those portions by mixing the signals at different levels of amplitude. All modifications were performed in Praat (Version 4.5.17, Boersma & Weenink, 2005). The syllables chosen were split in two at the boundary between the fricative and the vowel as determined by the onset of the first clear glottal pulse. The two portions of frication were trimmed to equal length by excising four 8 ms portions at 20% intervals of the total length. The vowel portions were equated in length, pitch, and RMS amplitude using Praat's manipulation object, which uses the Pitch Synchronous Overlap Add (PSOLA) method to align pitch periods, first equating duration, then pitch, both to an intermediate value between the two original recordings. Finally, the endpoint fricatives were interpolated to create a ten-step fricative continuum from one place of articulation to the other. This interpolation was done by adding up the signals at different ratios of amplitude, from a 0 retroflex – 9 alveolopalatal ratio for the alveopal end, to 9 retroflex – 0 alveolopalatal for the retroflex end. The same manipulation was performed on the vocalic portion. Spectrograms of the 2 endpoints for the frication and those for the vocalic portion are shown in Fig. 3.

The 10 fricative and 10 vocalic steps were orthogonally combined with one another yielding a bidimensional grid of 100 tokens as represented in Fig. 4. Twelve of these combination syllables were reserved for testing and the remaining 88 were presented during the initial exposure. During both phases, tokens were separated from one another by a 500 ms silence. In order to refer to the tokens in each trial, syllable combinations are denoted by referring to the steps in the continua where their components are located. For example, the syllable f6v0 refers to the combination of the fricative portion f6, generated by adding together the frication of the retroflex and the alveolopalatal tokens at a 6–3 ratio.
ratio of amplitude, with the vocalic portion v0, generated by adding together the vocalic portions of the retroflex and the alveolopalatal tokens at a 0–9 ratio of amplitude.

Notice that this type of interpolation does not reduce the complexity of the sounds and that multiple cues may be present in each dimension (e.g. both the pole frequency and overall spectral shape will vary in the fricative). An important consideration is how this manipulation affects the perception of the cues in the vocalic portion. In particular, this interpolation method produces intermediate tokens that essentially have doubled formants, one from each signal. Formants that are sufficiently close will be perceptually integrated (the center-of-gravity effect; see e.g., Chistovich & Lublinskaja, 1979; Delattre, Liberman, Cooper, & Gerstman, 1952; Xu, Jacewicz, Feth, & Krisnamurthy, 2004); specifically, listeners perceive a weighted average when formants are within 3–3.5 Bark. This effect is likely to be at work in average ratings of naturalness/unnaturalness on a scale from 1 (very unnatural) to 9 (very natural). Presentation was blocked by stimulus type, and the order of presentation was counterbalanced across listeners. Average ratings were not significantly different [t(7) = .75, p > .48; M = 7 (Maye) and 6.6 (ours)]. Additionally, to assess whether the manipulation in intermediate stages of mixing (resulting in doubled formants, as mentioned above) affected naturalness, an additional 8 naive listeners (3 English, 2 Italian, 2 French, 1 Swiss German) were asked to rate the endpoints (f0v0 and f9v9) and midpoints (f4v5 and f5v4) on the same 9-step naturalness scale, with the stimuli being presented in a randomized order. There was no difference in average ratings [t(7) = .73, p > .49; M = 5.13 (endpoints) and 4.38 (midpoints)].

2.1.3. Procedure and equipment

The experiment consisted of three phases, an initial exposure phase, a brief training phase, and a test phase. During the initial exposure phase, infants heard the familiarization tokens in a randomized, fixed order, while sitting on their caregiver’s lap in a small room. Infants across conditions heard the same total number of tokens (a total of 176 syllables; total presentation time was 157 s), and the same selection of tokens, but the frequency of specific tokens varied across the two distribution conditions, as represented in Fig. 5. Specifically, in the Flat distribution, infants heard every familiarization token twice, such that no combination of fricative and vowel was more frequent than other combinations. This condition represents a perceptual baseline, since infants are exposed to the same sounds but there is no mode in frequency to shape their perceptual space. In contrast, the Two Peak distribution suggested categories in the two natural endpoints, such that combinations of fricatives and vowels corresponding to the same category were more frequent. In order to keep infants’ attention and minimize distress, the auditory stimuli were synchronized with a visual display generated using the iTunes 6 viewer projected onto a large screen.

Immediately after the initial exposure, caregiver and infant crossed the hall to a testing booth. This booth consisted of a 3-walled enclosure made of white pegboard panels, approximately 4.5 ft high, with white curtains that descended from the ceiling to meet the pegboard. The pegboard was backed by thick cardboard to cover the holes, except for one large and two smaller openings in the front panel. The larger opening allowed a camera to record the session. A smaller opening allowed the experimenter to view the infant’s headturns. Finally, a third opening allowed...
a secondary observer, such as a second caregiver, to view the procedure. Both the experimenter and the caregiver holding the infant wore tight-fitting Peltor Aviation headphones through which they listened to loud masking music superimposed over low-level white noise. A chair was placed in the center of the booth, facing the front panel. A light was attached at the center of each panel, at the approximate eye level of an infant seated on a caregiver’s lap in the chair. The light on the front panel was green, while the lights on the side panels were red. Directly behind each red light, there was a Cambridge Ensemble II speaker. A Macintosh G4 computer fed the audio signal through a Yamaha Natural Sound Stereo Receiver RX-49 audio amplifier to these speakers.

Test trials began with the green light at the front flashing. When the infant oriented towards this light, it was extinguished and one of the red side lights began flashing. When the infant oriented towards the flashing side light with a 90° head-turn, the trial began. During any given trial, one pair of test stimuli was played through only one of the speakers, at the same time as the corresponding light was blinking. The stimuli continued to be presented randomly by the computer program used to run the study. In this testing booth, the experimenter cued the infant’s orientation towards the lights (and sound source) by means of a button box. The dependent measure was the amount of time that the infant oriented towards the light in each trial (Looking Time, LT).

Upon entering this test booth, infants first heard two training trials. In these first two trials, a maximum of 10 s of instrumental music was presented in order to introduce infants to the fact that sound presentation was contingent on their looking at the blinking light. After this brief training, infants were presented with the test trials proper, which reflect the combination of three variables: Dimension, Place, and Trial Type.

In order to keep the testing phase relatively short, each infant was tested on trials varying along only one Dimension. Half the infants were tested with trials in which tokens had the same fricative portion but the vocalic portion varied, and for the other half the vocalic portion remained constant and the fricative varied. In the first block of trials, the value of the unvarying portion was set to one place of articulation (e.g., retroflex), and in the second block of trials the value was set to the other place (e.g., alveolopalatal). The order of presentation of Place was counterbalanced across participants.

Within each block, 4 types of trials were presented, all of which consisted in 2 alternating tokens. In long trials, the tokens presented were the extremes along a single dimension and place (e.g., in the alveolopalatal, fricative-varying side, f0v0 and f9v0). This trial type allowed us to determine what drove infants’ preference with the present stimuli and procedure. Briefly, if infants were responding to acoustic distance between the two tokens being presented in a trial, looking times to the long trials ought to be maximal (if infants exhibited a preference for distinct tokens) or minimal (if they preferred similar-sounding tokens), given that these two tokens span the largest distance (both in interpolation steps and in acoustic terms).

The other three pairs spanned the same distance in terms of interpolation steps: we call these natural, mid, and unnatural. Mid trials consisted of the two tokens in the middle of one side (e.g., f3v0 and f6v0 for the alveolopalatal, fricative side; that is, when fricative varies, and the non-varying vocalic portion cues alveolopalatal place). The natural trials consisted of the token in a natural corner (that is, f0v0 – an alveolopalatal fricative combined with an alveolopalatal vocalic portion or f9v9 – a retroflex fricative combined with a retroflex vocalic portion) and the token closest to it among the ones reserved for test, three steps away along either the vocalic or the fricative dimension. Likewise, the unnatural pairs consisted of a token in the unnatural corner (f0v9 or f9v0) and the test token closest to it along either dimension.

2.2. Results

Before carrying out analyses, comparison of the long trials with the other three types allowed to determine the interpretation of the LT measure. If infants were responding to the perceptual distance between the two tokens being presented in a given trial, LT to long trials ought to be either maximal or minimal. Since this was not the case, as shown in Table 2, these trials were dropped from the analyses, as the interpretation of looking times is that of preference rather than discriminability. (A separate set of statistics confirmed that all factors and interactions found significant in the reported analyses remain so in analyses including these trials.) Thus, infants are likely not responding to the distance spanned between the 2 different tokens within a given trial. Looking time instead depended on the general acoustic characteristics of the 2 tokens presented. For example, looking times to the natural retroflex trials are best interpreted as those responding to the natural retroflex area of acoustic space, depicted on the bottom right corner in Fig. 4; looking times during the unnatural retroflex vocalic trials are due to tokens in the top right corner in the same figure, and so forth.

A repeated measures ANOVA with Distribution (Flat, Two Peak) and Dimension (Frication, Vowel) as across-subject factors, and Place (Alveolopalatal, Retroflex) and Trial Type (Natural, Mid, Unnatural) as within-subject factors on looking times as dependent measure revealed a significant 3-way interaction between Distribution, Place and Trial Type (F[2, 120]=5.17, p=0.007), and a significant two-way interaction between Trial Type and Place (F[2, 120]=5.88, p=0.004), and no other effects or interactions [all F values < 2, p > .16]. Looking times by Place and Distribution are shown on Fig. 6, to which we refer in the interpretation of the interactions. Notice that this Figure also shows the looking times in Experiment 2, which are presented and discussed in Section 3. As a reminder, the main factor of interest is Distribution; Dimension refers to which dimension was varying during test; Place to which category was being tested; and the Type to combinations of consonantal and vocalic portions (natural are those corresponding to the category, which therefore belong together and are highly frequent in the Two Peak condition; unnatural are infrequent and not instantiated in natural language.) The Trial Type by Place interaction emerges because infants’ looking times were consistently longer for the natural type within the retroflex place, but no such evidence was apparent in the alveolopalatal place. That is, infants’ LT were numerically larger for the natural combination of retroflex–consonant and retroflex–vowel than the unnatural combinations (retroflex–consonant and alveolopalatal–vowel or

---

**Table 2**

<table>
<thead>
<tr>
<th>Place</th>
<th>Long (Mean)</th>
<th>Natural (Mean)</th>
<th>Mid (Mean)</th>
<th>Unnatural (Mean)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Experiment 1: Flat</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Retroflex</td>
<td>11.3 (1.1)</td>
<td>13.0 (0.8)</td>
<td>10.4 (1.0)</td>
<td>11.7 (1.0)</td>
</tr>
<tr>
<td>Alveolopalatal</td>
<td>12.3 (0.9)</td>
<td>10.4 (1.0)</td>
<td>11.9 (1.1)</td>
<td>10.5 (1.0)</td>
</tr>
<tr>
<td><strong>Experiment 1: Two Peak</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Retroflex</td>
<td>11.9 (1.1)</td>
<td>13.0 (1.0)</td>
<td>10.9 (1.1)</td>
<td>9.0 (1.1)</td>
</tr>
<tr>
<td>Alveolopalatal</td>
<td>10.9 (1.1)</td>
<td>11.3 (1.1)</td>
<td>10.8 (1.1)</td>
<td>12.2 (1.0)</td>
</tr>
</tbody>
</table>
vice versa). However, they displayed no such base preference for natural alveolopalatal–consonant and alveolopalatal–vowel combinations over the unnatural combinations.

The 3-way interaction was further probed with a follow-up ANOVA within each place, in order to identify differences in perception as a function of initial exposure within the same regions of acoustic space. These follow-ups confirm that the 3-way interaction in the present experiment arises because infants’ looking times to the different trial types did not diverge depending on the familiarization distribution within the alveolopalatal place (that is, perception was not affected by the different frequency distributions; the follow-up ANOVA within the alveolopalatal place revealed no main effects or interactions [p > .05 for all]); in contrast, looking times did vary depending on the initial exposure within the retroflex place: While infants in the Flat distribution simply show an overall preference for natural trials, those that have heard a Peak distribution show a graded preference. This is evident by comparing the Flat and Two Peak looking times for the alveolopalatal place (on the right panel in Fig. 6, showing no clear patterns of preference), with those in the same conditions but in the retroflex place (left panel in Fig. 6, with statistically longer looking times to natural trials in both conditions, but a graded preference pattern only in the Two Peak Distribution condition). The follow-up ANOVA within the retroflex place reveals a significant effect of Trial Type [F(2, 60) = 8.4, p < .001; due to the preference for natural trials], as well as an interaction Type × Distribution [F(2, 120) = 3.08, p < .05; all other ps > .05]. The interaction Type × Distribution was investigated through post-hoc analyses, with the alpha set at .016, for 3 comparisons within each distribution. Infants who had heard a Flat distribution appeared to display some preference for the extremes of the space, although not significantly when controlling for multiple comparisons [natural-mid: t(31) = 2.42, p > .016; unnatural-natural: t(31) = 1.5, p > .016; mid-unnatural: t(31) = 1.25, p > .016]. In contrast, after hearing a distribution with two peaks infants treat natural and unnatural trials differently [t(31) = 3.83, p < .001], although there are no significant differences between natural and mid [t(31) = 1.98, p > .016], and mid and unnatural [t(31) = 2.18, p > .016]. In sum, infants’ perception was affected by the familiarization distribution with retroflex tokens, but was not affected with alveolopalatal ones.

2.3. Discussion

The goal of this experiment was to extend previous findings regarding infants’ ability to learn categories from distributions in the acoustic signal in two ways. First, two non-salient categories were chosen; and second, two dimensions were varied simultaneously during the initial exposure, although only one was informative during test, to assess whether infants were tracking both dimensions during the learning phase. As for non-salience, results suggest that statistical learning may extend to some, but not all less salient categories. Indeed, infants’ perception of natural and unnatural retroflex consonant–vowel combinations was different in the baseline as compared to the Two Peak Distribution condition, suggesting some reorganization of perceptual space around the retroflex prototype after training – in other words, infants may have learned the retroflex category. However, no learning was evident in the alveolopalatal place of articulation, with no differences as a function of exposure. This difference in learning across the two places of articulation was supported by a significant interaction between test Trial Type, Place of articulation, and Exposure Distribution, and confirmed through post-hoc comparisons. With respect to multidimensionality, since infants showed different looking times to the different Trial Types in at least some of the experimental conditions, it would appear that they could discriminate tokens on the basis of limited cues (those available in the consonant, or the vowel, depending on the infant). Furthermore, the lack of interactions with dimension is consistent with the hypothesis that infants were tracking both dimensions during the initial exposure and could rely on limited cues during testing.

As mentioned in the Introduction, perceptual acquisition of non-salient sounds may involve enhancement, where initially weak abilities are tuned, or induction, where there is no evidence of initial abilities. As summarized in Section 1.3.1, while both bilsibans used in the present study are non-salient, some evidence suggests that retroflexes are not inherently weak (see also Hamann, 2005, for further evidence to this effect). Additionally, looking times to natural retroflex combinations were longer than less natural combinations after the simple exposure to the Flat distribution. This may indicate that infants are sensitive to the well-formedness of this combination even in the absence of experience with the distributions encountered in natural languages; more importantly, this indicates that infants start out with a fine-grained perceptual sensitivity to distinctions within those regions of acoustic space. In the Two Peak condition, repeated presentation of tokens near natural retroflex combinations emphasized (dis)preferences that were partially present after a flat exposure, resulting in long looking times towards natural retroflex combinations and shorter ones to the unnatural retroflex combinations. Thus, the case of retroflex place discrimination appears to be an example of enhancement: initially weak sensitivities are ameliorated by exposure to a bimodal statistical distribution.
Alveolopalatals present a different case, since no such baseline preference for the alveolopalatal natural combinations were evident after infants were exposed to the flat distribution. Interestingly, no learning of the natural cues in this area of perceptual space appears to have taken place in infants exposed to the structured Two Peaks distribution either. In this context, one can liken the situation to that of negative elasticity, that is, absence of perceptual learning in the absence of prior sensitivities. However, there is an alternative explanation that ought to be ruled out before entertaining this possibility. One confounding factor in the present experiment was that infants not only had to learn a non-salient category in the absence of obvious pre-existing abilities; but also that initial exposure included a competing retroflex category, for which infants appear to have some predilection. More generally, if 2 categories occur in the input, in the presence of limited resources, only the more salient one may be learned.

As discussed in Section 1.3.1, there is some evidence suggesting that the retroflex sibilant tends to dominate perceptual judgments, since Polish and Mandarin adult listeners’ identification of alveolopalatals and retroflexes is asymmetrical. Furthermore, the potential dominance of the retroflex place category has been documented with the stimuli used in the present experiment, which are drawn from McGuire (2007). There, it was reported that listeners whose language background contains both alveolopalatal and retroflex sibilants tested on the same stimuli used in Experiment 1 tend to allocate a more restricted area in acoustic space to the alveolopalatal category (see, Figure 2.4 on p. 42 for Polish, and Figure 3.7 on p. 74 for Mandarin, both in McGuire, 2007). Even listeners whose ambient language does not contain these two categories tend to label the retroflex end of the grid more consistently than the alveolopalatal one (see Figure 3.6 on p.73 of McGuire, 2007).

Thus, it is still possible that infants could learn the alveolopalatal category in the present setup provided that there were no competition from a similar category. A second experiment was carried out to assess this possibility, by presenting infants with numerous tokens around the alveolopalatal natural category only. We predicted that, if infants succeeded in learning this sound, a significant effect of Trial Type should be found, such that, similarly to that encountered for the retroflexes, infants would exhibit a preference for the natural combinations of alveolopalatal frication and vocalic portions.

3. Experiment 2: alveolopalatal category

In this experiment, infants heard tokens near the alveolopalatal natural combinations much more frequently than any other combination. While no changes were expected in the retroflex end of the grid, the design was the same as that in Experiment 1, in order to allow for a comparison of performance with infants in the flat condition of Experiment 1.

3.1. Methods

3.1.1. Participants

Thirty-two English monolingual, fullterm infants were included ($M_{age}=4.99$ months, range 4.18–5.82 months, 21 female). The same exclusionary criteria as in Experiment 1 were applied here, resulting in 8 additional infants not included for the following reasons: being exposed to a language other than English (7); or having looking times shorter than 1 second on any given trial (1).

A repeated measures ANOVA within the alveolopalatal place with Dimension (Frication, Vowel) and Type (natural, mid, unnatural) revealed no main effects or interactions [all Fs < 1]. Furthermore, an ANOVA including the flat condition revealed no effect of distribution nor interactions with that variable, suggesting that performance in the alveolopalatal place was not statistically different after a flat familiarization and after a familiarization with a highly kurtotic unimodal distribution with its mode centered over the alveolopalatal end of the grid. Looking times for this experiment are reported in Table 3.

Thus, infants failed to show any significant preference within the block where the constant dimension was consistent with an alveolopalatal place of articulation, even after hearing many more repetitions of alveolopalatal tokens. Since the exposure distribution infants heard in this study did not emphasize the more salient retroflex, infants’ failure in the present study cannot be attributed to their attending only to the competing retroflex category. By extension, it is unlikely that this was the reason why infants failed to learn the alveolopalatal combinations in the Two Peak condition in Experiment 1.

Furthermore, notice that the proportion of repetitions used was much higher than in the Two Peak condition (which is 13 times as many as that of infrequent tokens) and than in previous work (e.g., 4 times as many as that of infrequent tokens in Maye et al., 2002, 2008). Nonetheless, it is still possible that our training simply was not long or intense enough. In order to support this explanation, it would be necessary to carry out additional experiments increasing the length of exposure until a non-null result was found. Since the process of lengthening the exposure within...
an experimental setting could continue ad infinitum, we leave that endeavor for future research, and advance the provisional conclusion that some non-salient categories are more amiable to learning than others.

4. General discussion

Previous research (Maye et al., 2002, 2008) has provided convincing evidence that infants’ sensitivities can be shaped even by brief exposure to the distribution of a single, salient acoustic cue in a simplified perceptual space. The goal of the present study was to assess the effects of exposure to different distributions of acoustic cue in infants’ perception of a pair of non-salient sibilants, which were cued by multiple dimensions. Results suggest that multidimensionality did not pose a problem, while different learning outcomes ensued for the two sibilants. Each of these findings are discussed in more detail in the next 2 subsections, and their theoretical implications are drawn out in Section 4.3.

4.1. Extension to non-salient categories

In the present study, infants’ perception of 2 sibilants differing in place of articulation was assessed after different exposures to a bidimensional grid across their vocalic and frication cues. In a condition that acts as a baseline, infants were simply exposed to the whole grid with equal frequency. After this exposure, they exhibited somewhat longer looking times to natural retroflex combinations, while no such preference appeared in the alveopalatal trials. In a second condition, infants heard many more tokens in the acoustic area corresponding to natural retroflex and natural alveopalatal categories. After this experience, infants’ preference for natural retroflexes and dispreference for unnatural retroflex combinations reached statistical significance, a result compatible with a reorganization of perceptual space triggered by learning through exposure to acoustic cue distributions. In stark contrast, no such learning occurred for the alveopalatal series. In view of the repeatedly documented fact that sibilants are challenging for infants (see Section 1.3.1), the retroflex results support the hypothesis that infants can learn some non-salient categories by relying on frequency distributions in acoustic space, while the difference between retroflexes and alveopalatals in the baseline and the experimental conditions may shed light on additional effects of salience.

Indeed, pre-existing sensitivities involving retroflex tokens may have enabled infants to attend to the frequency distributions in this region of acoustic space, thus constituting a necessary condition for learning. That is, infants’ attention to natural retroflex combinations could have acted as an anchor for the distributions encountered in the input. Given the lack of preferences for the alveopalatal tokens in the baseline condition, no such perceptual bootstrapping could happen for the alveopalatal sibilants. It may not be by chance, then, that there was no evidence of learning after exposure to alveopalatal tokens. We return to ways in which this situation may be resolved in Section 4.3.

In short, the present study extends the effects of exposure to acoustic cue distributions beyond the realm of VOT. These results strengthen the power of the statistical learning explanation to encompass dimensions that have not been documented as being (psychoacoustically) salient. At the same time, they do not rule out the possibility that statistical distributions are not a sufficient condition for learning to take place, but instead some minimum of sensitivity to the acoustic dimensions involved may be a necessary condition instead.

4.2. Extensions to multi-cue contrasts

While a great deal of work has investigated the effect of multidimensionality on adult perceptual learning, the infant literature has lagged behind, with only a few studies documenting infants’ discrimination abilities in the presence of limited cues. In this context, the present study provides a very first insight, by testing infants’ ability to learn categories based on multiple varying acoustic correlates.

Specifically, the alveopalatal and retroflex categories were cued through varying frication and vocalic dimensions during the initial exposure. Naturally, in a multidimensional grid where dimensions are well correlated, as in the present case, listeners are still free to attend to only one dimension. Therefore, to ensure that infants could not succeed by attending to a single correlate in a single dimension, only one dimension varied and the other was rendered uninformative by keeping it constant throughout the test. Given that infants’ looking times to the test trials varied depending on the initial exposure, which spanned both dimensions, this suggests that even though they could have attended to a single dimension, they kept track of both dimensions, and could later resolve the task on the basis of correlates within either dimension.

These results constitute a limited, initial step in approaching multidimensional category learning in infancy, and they necessarily leave open a myriad of questions that have occupied the field of adult category learning. It may be useful to point out three specific questions related to multidimensional speech categories that must await further research. First, we mentioned that there were no interactions with the factor Dimension, suggesting that performance was not markedly worse when either the vocalic or the frication dimension was unavailable. Naturally, this may have been due to ceiling or floor effects, and it does not preclude that in a different testing situation there may be a difference between infants’ ability to rely on the 2 types of information. In other words, the absence of a difference found here cannot be interpreted as evidence of absence of a weighting bias. Secondly, it remains an open question whether infants, like adults, perceptually integrate co-occurring cue values to the point that discrimination of tokens containing conflicting cues is markedly worse than stimuli containing correlated cues (but, as mentioned, Ko et al., 2009 document that by 14 months infants prefer stimuli containing cues that are correlated in the ambient language). Finally, as summarized above, multidimensional categories appear to be harder for adults than unidimensional ones in speech. Furthermore, this is also the case for adult learning of non-speech and non-auditory categories, and other work suggests a similar bias in non-human animals, all evidence converging towards unidimensional categorization being the default. Given that we did not compare unidimensional and multidimensional categories, our data cannot contribute to this question directly. However, we summarize in Section 4.3 some work suggesting that a radically different type of multidimensionality, namely multimodality, actually boosts infants’ learning.
4.3. Implications for theories of infant phonetic acquisition

As explained in the Introduction, the statistical learning hypothesis may provide a parsimonious and comprehensive explanation for the multiple processes involved in the developmental changes in phonetic perception documented in the first year of life. Previous laboratory-training studies have begun to provide empirical support to this hypothesis, by showing that infants’ perception of VOT is shaped by the distributional properties of the input they are exposed to. In particular, Maye and colleagues have documented 2 of the 4 postulated processes, maintenance (Maye et al., 2002, 2008, bimodal condition), and attenuation (Maye et al., 2002, 2008, unimodal conditions), by testing category learning along a dimension that is psychoacoustically salient.

An equally important test of the statistical learning hypothesis is whether it can accommodate learning of initially weak, but nonetheless existing, sensitivities. The case for retroflexes in the present paper provides this evidence, extending previous findings to a third learning path, that of enhancement.

In contrast, the last type of statistical learning remains elusive: Does pure induction based on acoustic cue distributions ever take place? In this paper, we found that for one non-salient category, there was no evidence of prior sensitivity, and no evidence of statistical learning; thus, we could not document induction in the laboratory. One may argue that the developmental pattern documented for [n] and [ŋ] provides evidence for acoustically based induction, as younger infants fail entirely to make this distinction, while 12-month-olds succeed (Narayan et al., 2009). However, while at this age infants probably cannot make use of top-down lexical knowledge, they do have access to information other than acoustic cue distributions, as the acoustic signal can be yoked to visual information for visible articulators, and to proprioceptive information for the sounds that infants babble. Indeed, visual cues are available for the [n-ŋ] distinction (Johnson, DiCanio, & MacKenzie, 2007), and most infants babble velar nasal consonants (e.g., Locke, 1983; Robb & Bleile, 1994), such that both types of information may underlie the acoustic cue distributions associated with these categories. The effect of visual cues on perceptual acquisition has been highlighted by work comparing confusion matrices in blind versus seeing children (Mills, 1987); and research on adult second language acquisition (Hardison, 2003).

Instead, more appropriate evidence would come from cases like [d] and [ð] (Polka et al., 2001), and [s] and [ʃ] (Nittouer, 2001), where visual cues are not robust (see Babel & McGuire, 2010 for [ð]) and proprioceptive information is not available to guide infants’ attention towards non-salient acoustic cues. Interestingly, neither of these cases is resolved by 12 months, lending indirect support for the possibility that at least some non-salient categories require the conjunction of multiple types of information in order to be learned. The possibility that infants should rely on multiple types of information to learn acoustically fragile categories is not in opposition to the idea that multidimensional categories are harder than unidimensional ones, which appears to be the case for adults (and possibly non-human animals; see Section 1.3.2). On the contrary, there is an important difference between multidimensionality in a single modality, and having access to multiple correlated cues in different modalities, which can contribute to heighten attention and indirectly improve performance (Bahrick & Lickliter, 2000). Previous research suggests that infants benefit from intersensory redundancy when learning perceptual rhythmic categories (Gogate & Bahrick, 1998), abstract patterns (Frank, Stemmer, Marcus, & Johnson, 2009), and sequential order (e.g., Lewkowicz, 2004), finding it easier to learn, generalize, and discriminate when multiple senses are involved.

5. Conclusions

In short, previous laboratory-training studies replicate natural language patterns of maintenance and decline: initially robust sensitivities are shaped by frequency distributions along a single salient correlate. The present study explored the power of statistical learning on the basis of distributions of a multi-cue contrast to a case of enhancement: Non-null prior sensitivities are improved by acoustic cue distributions. We also document a case of inelasticity, where poor sensitivities are not shaped by acoustic cue distributions; although they remain unpublished, laboratory-based replications of the other type of inelasticity (high sensitivity not affected by acoustic cue distributions) have been reported at conferences (Pons, Migitani, Amano, & Werker, 2006; Pons, Sabourin, Cady, & Werker, 2006). In contrast, it is unclear whether there are laboratory-based or natural language acquisition data documenting induction on the basis of purely acoustical evidence. Indeed, information from other senses (visual and proprioceptive) and even rudimentary lexical categories (Swingley, 2009) may be necessary, or at least useful, to learn fragile acoustical categories. Overall, these results extend the claim that the diversity of paths to phonetic acquisition in the first year of life. While acoustic cue distributions certainly contribute to re-shaping perceptual space into categories even in early infancy, future work should further explore a quantification of this impact, its interactions with baseline auditory-perceptual abilities, and with other sources of information.

Acknowledgments

This work was supported by funds from Purdue University to AC and AS; from Ecole de Neurosciences de Paris and Fondation Fyssen to AC; from NIDCD R01 DC004421 (Keith Johnson) and funds from UC Santa Cruz to GM; from NICHD R03 HD046463-0 to AS; from NIH R03 DC006811 to AF. The authors are particularly grateful to the anonymous reviewers and Cassie Mayo, whose comments substantially improved this manuscript. Portions of this work were presented in 2008 at the Acoustical Society of America November Meeting, the American Speech-Language-Hearing Association Meeting, and the International Child Phonology Conference.

References
