Perceptual attunement in vowels: A meta-analysis

Sho Tsuji[1,2] and Alejandrina Cristia[3,4*]

[1] Radboud University

[2] International Max Planck Research School for Language Sciences

[3] Laboratoire de Sciences Cognitives et Psycholinguistique, CNRS, ENS-DEC-EHESS

[4] Neurobiology of Language, Max Planck Institute for Psycholinguistics

* Corresponding author: Alejandrina Cristia, 29, rue d'Ulm, 75005, Paris, France.
alecristia@gmail.com

Abstract

Although the majority of evidence on perceptual narrowing in speech sounds is based on consonants, most models of infant speech perception generalize these findings to vowels, assuming that vowel perception improves for vowel sounds that are present in the infant's native language within the first year of life, and deteriorates for non-native vowel sounds over the same period of time. The present meta-analysis contributes to assessing to what extent these descriptions are accurate in the first comprehensive quantitative meta-analysis of perceptual narrowing in infant vowel discrimination, including results from behavioral, electrophysiological, and neuroimaging methods applied to infants 0-14 months of age. An analysis of effect sizes for native and non-native vowel discrimination over the first year of life revealed that they changed with age in opposite directions, being significant by about 6 months of age.

Keywords: Development; humans; infancy; language; meta-analysis; speech; vowels

Infant vowel attunement: A meta-analysis

1 Introduction

Over the last 50 years, the experimental study of infant speech sound discrimination has provided us with important insights into early perceptual abilities and their change as a function of development and language exposure. Much attention has been paid to perceptual narrowing: Infants are thought to start out with language-universal perceptual abilities (i.e., patterns of perception that are independent of language exposure), and these abilities would become tuned to the infant's ambient language as a function of exposure, culminating in the end of the first year of life with qualitatively different patterns of perception by infants exposed to different languages.

Perceptual narrowing provides crucial insights on the psychobiological bases of language because it is the first sign that infants are acquiring their native language. Therefore, attunement can shed light on the complex interplay of biological and experiential factors involved in the unfolding of linguistic abilities. For instance, we have recently learned that infants exposed to serotonin reuptake inhibitors prenatally show perceptual attunement *earlier* than control infants (Weikum, Oberlander, Hensch, & Werker, 2012). Additionally, individual variation in attunement predicts later language development (a recent review in Cristia et al., in press). Compared to consonants, vowels are more clearly heard in the womb (a recent summary in Granier-Deferre, Ribeiro, Jacquet, & Bassereau, 2011). Therefore, attunement for vowels results from speech exposure starting even before birth, and it has been thought to be evident earlier than consonants (a question we revisit below). Thus, vowel discrimination scores could be

particularly useful to make decisions regarding both the at-risk status of specific infants and their priority for treatment, and the short-term effects of early treatments, at a very young age.

An additional reason for studying perceptual narrowing in vowels is internal to the field of infant speech perception. In fact, the majority of evidence for perceptual narrowing in speech perception comes from consonants. Nevertheless, prominent models of early speech perception by and large consider perceptual narrowing to apply to all speech sounds rather than to consonants in particular. Therefore, it is crucial to assess how far such generalization is suitable, as some evidence suggests that vowels and consonants are not completely comparable. To begin with, a host of infant, child, and adult psycholinguistic evidence suggests that they are not processed in precisely the same way (e.g., Bonatti, Peña, Nespor, & Mehler, 2004; Caramazza, Chialant, Capasso, & Miceli, 2000 and references therein). Moreover, while infants' perception can change with brief lab-based exposures to consonants (e.g., Cristia, McGuire, Seidl, & Francis, 2011 and references therein) and lexical tones (Liu & Kager, 2011), such perceptual warping has failed to occur for vowels (Pons, Sabourin, Cady, & Werker, 2006; Pons, Mugitani, Amano, & Werker, 2006). Based on these substantial differences in findings on vowels and consonants, it is of particular interest to revisit the question of perceptual narrowing for vowels specifically.

Before turning to the quantitative study, we will provide a brief overview of a few prominent models of perceptual narrowing in infant speech perception. The Native Language Magnet model (NLM; Kuhl, 1994; Kuhl, Conboy, Coffey-Corina, Padden, Rivera-Gaxiola, & Nelson, 2008) was originally based on evidence from vowel

discrimination (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992), and it is better specified than the others models in terms of when and how vowel perception becomes attuned to the native language (e.g., Kuhl et al., 2008). For this reason, we expand on this particular model and the evidence supporting it first.

The perceptual magnet effect refers to the phenomenon that vowel tokens are treated differently depending on how prototypical they are of a vowel category. Vowel prototypes in the context of NLM have been described as the representations most often activated (Kuhl et al., 2008), or as the centers of a vowel category (cf. Feldman, Griffiths, & Morgan, 2009). With exposure to the native language, prototypical vowels start acting like magnets, warping perceptual space such that it shrinks around prototypical vowels and creates non-linearities in perception. Thus, discrimination of tokens close to a prototype becomes worse than discrimination of tokens towards the category boundary. Since warping depends on exposure to sounds mapping on native vowels, no such magnet effect occurs for non-native vowels.

Early evidence for language-specific vowel perception relied on non-linearities in the detection of within-category changes. A first indication for native vowel prototypes was given in two studies on 6-month-old English-learning infants, who were better able to discriminate vowels in the direction from a non-prototypical to a prototypical native exemplar of [i] (the vowel in the word 'sheep') than vice versa (Grieser & Kuhl, 1989; Kuhl, 1991). The seminal Kuhl et al. (1992) subsequently documented that American English 6-month-olds failed to detect many vowel changes around the prototypical [i] in their language but were sensitive to the same acoustic distances centered around [y], while Swedish infants tested with the same stimuli readily heard such changes around the

non-native [i] and missed them around native [y]. Based on this evidence, Kuhl and colleagues proposed that narrowing occurs earlier in vowels (by around 6 months) than in consonants (closer to 8-10 or as late as 10-12 months; Werker & Tees, 1984). The NLM model in its current form is not restricted to within-category changes, and has been invoked in several studies that document developmental changes (Polka & Werker, 1994), cross-linguistic differences (e.g., Bosch & Sebastián-Gallés, 2003), or cross-contrast differences (better discrimination for a native than a non-native contrast, e.g., Cheour et al., 1998; but see Best, McRoberts, LaFleur, & Silver-Isenstadt, 1995).

NLM is not the only model that has been put forward to account for infant speech processing, and could thus capture the aforementioned changes in vowel discrimination. The Perceptual Assimilation Model (PAM; Best, 1994) is also well known. However, it provides an account primarily in terms of how non-native sounds are processed once native perceptual categories have already been formed, rather than explaining the process by which native and non-native categories come to be treated differently, and thus it is not a model of perceptual attunement. We note here that PAM will become relevant once more in the final discussion below.

The developmental framework for Processing Rich Information from Multi-dimensional Interactive Representations (PRIMIR; Werker & Curtin, 2005) is another mainstream model of infant speech perception. In this model, perception always must be conceived as operating in multiple levels or planes at the same time. One of these is the General Perceptual plane, which encodes discrimination abilities that are initially independent of language exposure, and thus very similar in infants exposed to different languages. As a function of language experience, including not only listening but also

visual and articulatory experience, this plane is somewhat reorganized reflecting the native language categories, such that some innate boundaries are erased, enhanced, or shifted. This model also states that this representation, albeit language-specific, is not very robust or abstract. True phonological categories will only emerge as the child begins to learn words and store them in the Word Form plane, at which point a third plane (Phoneme plane) will begin to be developed (compare this with the Word Recognition and Phonetic Structure Acquisition, WRAPSA model, e.g. Jusczyk, 1993). Thus, PRIMIR differs from NLM in several aspects with regards to perceptual attunement. First, it more openly incorporates visual and articulatory experience in the process of attunement. Second, it predicts that reorganization may also be brought about by word learning.

Aside from these differences, both PRIMIR and NLM hold that infant vowel perception changes over the first year, with native discrimination improving and non-native discrimination deteriorating. As mentioned above, there is some evidence in favor of this view. However, other studies fail to find developmental changes (which are assumed to be due to experience) or cross-linguistic differences within the first year of life (e.g., Polka & Bohn, 1996; Sebastián-Gallés & Bosch, 2009). Moreover, where developmental changes are indeed reported, the timepoint of their occurrence is debated. While some studies find a modulation by 6-8 months of age (e.g., Bosch & Sebastián-Gallés, 2003; Kuhl et al., 1992; Polka & Werker, 1994), others only find modulations from 10 months of age onwards (e.g., Polka & Bohn, 2011; Pons et al., 2012). Therefore, based on these studies it is far from clear that the reorganization for vowels is truly robust; and that it happens earlier than 6 months.

Given the considerable diversity in outcomes, it was relevant to assess the evidence for perceptual narrowing in vowels critically. To this end, we carried out a comprehensive review of the vowel discrimination literature, and identified studies where two or more age groups of infants had been tested on the same vowel contrast. We then retrieved or calculated the effect size indicative of discrimination in each case, and combined effect sizes using meta-analytic methods, as explained in detail in the next section. We sought to answer the following questions. First, do effect sizes change differently with infant age depending on whether the contrast is native or non-native? A change in opposite directions for native and non-native contrasts and with a more positive slope for native contrasts is indicative of perceptual narrowing. Subsequent questions investigated specific features of this process: Second, does native contrast discrimination improve with age? Third, does non-native discrimination deteriorate with age? Finally, do these changes occur by about 6 months?

2. Methods

2.1 Search protocol

A full search on scholar.google.com was conducted in September 2012 with the keyword combination "{infant|infancy} & {vowel|speech sound|syllable} & discrimination". Additionally, the search terms were translated into French, German, Japanese, and Spanish for according searches. We also asked experts in the field to inform us of any published or unpublished studies we had missed. Experts were defined as scientists having participated in at least 2 studies identified in our intermediate search

sample or who were part of a lab where such research had taken place, and who were still active in the field or could be otherwise contacted. Further, articles were added based on a screening of articles cited and articles citing the articles in the remaining search sample. The complete sample is available as a public resource (Tsuji & Cristia, in preparation, https://sites.google.com/site/inphondb/).

The search sample was narrowed down to the final search sample of 21 articles based on the following inclusion criteria: (1) The study focused on normally developing infants, with at least one age group involved being 12 months of age or less. (2) At least two age groups were assessed on the same vowel contrast. (3) Discrimination was the key component of the task. (4) The two stimuli being discriminated were described as differing only in vowel quality or quantity. (5) The two stimuli being discriminated were auditory only. If a visual stimulus was presented, it was only for the purpose of indirectly measuring infants' attention by looking time, or in order to distract infants with unsystematic stimuli. (6) The articles was published in any source, including peer-reviewed journals (N = 15, in addition, 2 articles are under review: Benders, submitted and Mazuka, Hasegawa, & Tsuji, submitted, and 2 articles are in preparation: Liu & Kager, in preparation a, and Liu and Kager, in preparation b), conference proceedings (N = 1), and theses (N = 1). Given that the key question pertained to the first year, we excluded records focusing on infants older than 15 months of age.

The 21 articles of the final search sample contained 116 eligible records. We define a record as an experimental unit for which a separate result was reported. In most cases, this was one experiment on one group of infants, but sometimes it was the case that, for

instance, values for different orders of presentations were reported separately. In such cases, we counted each reported unit as one record.

2.2 Experimental methods for assessing infant speech sound discrimination

Before turning to the quantitative analysis, we will give a short overview of the methods used to assess speech sound discrimination in infants. Along with the methods themselves, we will outline the respective dependent variables on which later effect size calculations were based. Although the methods combined in this meta-analysis are varied, they all assess the same construct, namely infants' response to a sound change. As such, they are suitable for combination into one meta-analysis.

Central Fixation (CF), also sometimes referred to as Visual Habituation, is a paradigm where a central audiovisual stimulation is presented contingent on the infants' attention (for details, see Werker, Cohen, Lloyd, Casasola, & Stager, 1998). Therefore, it can be used in combination with habituation-dishabituation designs, where the same stimuli are presented repeatedly until attention wanes. It can also be used in familarization-preference designs, where the initial exposure is fixed in duration (rather than dependent on a decline of attention). In both cases, the habituation or familiarization phase is followed by a test phase, in which the infant is presented with one or multiple trials of the same stimulus, as well as one or multiple trials of a novel stimulus. The looking times to the same and to novel trials are the dependent variables, and the difference in looking times is assessed within-participants. All but one of the studies using CF in the current sample followed the above design. One study (Benders, submitted) employed the stimulus alternation design, a variant of CF in which infants are presented non-alternating

trials with repetitions of the same stimulus as well as alternating trials in which the same stimulus alternates with a novel stimulus, without a prior habituation or familiarization phase. The study with this design assessed differences in looking times by calculating the ratio of look duration during alternating trials divided by the look duration during the surrounding non-alternating trials.

In the Headturn Preference Paradigm (HPP), audiovisual stimulation is presented on the right and left sides of the infants contingent on their head-turns to the respective sides (for details, see Kemler Nelson et al., 1995). Like CF, HPP can be used in familarization-preference designs such that the infant is initially exposed to repetitions of the same stimulus until a fixed looking time has accumulated. In the subsequent test phase, the infant is presented with multiple trials of the same or a novel stimulus, which are presented on either the left or the right side paired with a flashing light in pseudo-random order. The difference in infants' orientation times to trials with the same or novel stimulus is assessed within-participants.

The Conditioned Head-Turn (CHT) paradigm also makes use of infants' headturns towards a visual reinforcement. Infants are trained to respond to sound changes by turning their head towards a visual reinforcement each time there is a sound change. At a subsequent stage, the visual reinforcement becomes conditional to correct headturns (details in e.g. Werker, Polka, & Pegg, 1997). After training infants on this contingency, they are tested on the sound contrast of interest (sometimes on several contrasts over subsequent days). A single measure per participant, such as the percent of correct headturns to a sound change is reported as the dependent measure. While some studies also report the sensitivity measures d-prime or a-prime, we base our effect size

calculations of percent correct in the current sample because this was the measure consistently reported in all studies.

In electroencephalography (EEG), the electrical activity of the brain is measured with electrodes placed on the scalp. Infant speech sound discrimination has often been measured through the mismatch response (MMR), an event-related potential (ERP) response that appears when a rare (deviant) stimulus is presented in a row of repeated (standard) (for details, refer to Cheour, Leppänen, & Kraus, 2000). As the method does not require attention to stimulation, infants are often silently entertained with toys or a silent movie during the experiment. The MMR is defined as the difference wave between the response to standard and deviant stimuli. Both the latency and amplitude of the MMR constitute important measures. For the purpose of the current study, we chose to base effect size calculations on the amplitudes. The auditory MMR in adults occurs as a fronto-central negative potential at around 150-250 ms after onset of stimulation, while in infants both positive and negative polarities in a broader time-range are observed. In one of the two EEG studies included in the final analysis, the MMR was defined as the most negative peak in a time window of 200-500 ms, and amplitude was calculated from a 50 ms time-window centered around the peak at right frontal electrode F4. In the other study, the MMR was defined as the most negative peak in a time-window from 150-300 ms, and amplitude was calculated as the average over fronto-central bilateral electrodes F3, C3, P3, F4, C4, P4 in a 100 ms time-window centered around the peak.

Near-infrared spectroscopy (NIRS) measures changes in hemoglobin oxygenation in specific brain regions. Speech sound discrimination in infants is measured by presenting blocks in which a single (type of) stimulus is repeated, as well as "alternating" blocks, in

which that stimulus is interspersed with a novel one. As in EEG, infants do not need to attend to stimulation and are often entertained with unrelated visual stimuli during the experiment. Two types of dependent variables have been typically used for measuring speech sound discrimination in infants: changes in oxygenated or deoxygenated hemoglobin concentration between the two types of blocks mostly in probes over the superior temporal gyrus (STG) in the left hemisphere, or a laterality index calculated from probes over STG in both hemispheres, indicating how selective the activation is. As the former is regarded as a measure of pure discrimination, while the latter is regarded to reflect more linguistic processing, we aimed to include the former in the analysis. However, for the three studies included in the final analysis, we succeeded in retrieving the former in two, and the latter in all three studies. We therefore decided to calculate the effect sizes based on the laterality index for all three studies.

We decided on the effect size measure by experimental method as outlined below. We then divided the articles randomly and coded them independently. After the coding process, records were cross-checked for inconsistencies several times.

2.3 Selection of samples and coding of effect size

Of the 116 records, we succeeded in calculating effect sizes for 100 records (86%) out of 17 studies (cf. Table 1 for an overview of studies for which effect sizes could be calculated). The articles of which we were able to calculate effect sizes were published between 1992 and 2012 (2 were under review and 2 in preparation) by 13 different first authors. Following standard meta-analytic practice, we removed outliers above or below 3 SD from the sample mean (Lipsey & Wilson, 2001). Three records were removed by

this criterion (cf. Fig.1). Thus, the final dataset included 97 records, 75 for native and 22 for non-native. The records were based on a total of 1613 unique infants, some of them measured repeatedly for a total of 1882 unique measurements.

---Insert Table 1 around here---

Effect sizes were calculated based on Lipsey and Wilson (2001). As outlined in 2.2, depending on the method, the outcome was either reported as a comparison between two conditions within one group of infants (CF[1], HPP), or a single score that could be a ratio (one CF study), a difference score (ERP, NIRS), or a percentage (CHT). Cohen's d, an effect size measure that involves dividing the differences in means by their standard deviation, was calculated in all cases. As the majority of records had a sample size < 20, Hedges' correction for small samples was applied to all effect sizes.

In CF and HPP studies (57 records), the difference between same and novel trials in the test phase was a within-subject measure. For these two methods, the standardized mean gain effect size for within-subject comparisons (Lipsey & Wilson, 2001) was calculated, in which the mean difference score between same and novel trials is divided by their pooled standard deviation. In calculating the standard error of the standardized mean gain effect size, the correlation between the means of the same and novel trials is taken into account. The inclusion of a correlation term leads to a smaller standard error the larger the correlation, thus taking into account the increased precision of within-subject measures. This correlation was not reported by any of the studies included, but we were able to obtain the original correlations from the first authors of six studies (personal

1! Excluding one study using the stimulus alternating paradigm and calculating a ratio as the outcome variable.

communication), which covered 42 experiments. For the remaining 15 experiments, we chose the median correlation of these 42 data points, which was $r = 0.505$ ($SD = 0.255$).

All other studies reported one value per record. This value could either be a ratio (one CF study, 3 records), a difference score (ERP and NIRS, 23 records), or a percentage (CHT, 14 records). For these cases, we calculated the standardized mean difference score (Lispey & Wilson, 2001) for between-subject comparisons. This effect size is equivalent to the standardized mean gain score when sample sizes of control group and experimental group are the same. In order to calculate the effect size, we assumed a control group performing at the respective chance level (1 for the CF study, 0 for ERP and NIRS, 50% for CHT). The standard error of the effect size for uncorrelated samples was calculated. The weight of all effect sizes was obtained as the inverse of the squared standard error.

2.5 Coding of moderator variables

The only relevant participant characteristic for the present analyses was infant age. We entered mean or median age in days into the analysis. If a range was reported instead of a mean or median, we chose the midpoint of the range as an estimator of age. If only age in months was reported, we estimated the age in days by multiplying the number of months by 30.42. We were able to estimate age for all experiments based on these procedures.

The only relevant stimulus characteristic included in the current analyses was the phonemic status of the stimulus in the infants' native language[2]. Stimuli were coded as

[2] Additionally, we coded measures of spectral and temporal distance between stimuli. Spectral distance refers to differences in vowel formant frequencies, and temporal distance refers to differences in vowel length. For the present sample, a spectral distance could be estimated for only 60% of records, and a temporal distance for 36% of records.

native if the vowels were reported to be present in the vowel inventory of the language by the authors. All other stimuli were coded as non-native. Non-native stimuli could thus either be non-native vowels, or speech sounds that were modified such that they were not contrastive in that the infants' native language. The latter was the case for two studies using a vowel length distinction outside of the contrastive range for the native language (e.g., Minagawa, Mori, Naoi, & Kojima, 2007), and one study where one of a pair of identifying features was neutralized (either quality or length, Benders, submitted).

3. Results

3.1 Preliminary Analyses

A set of preliminary analyses was conducted to assess overall sample characteristics. We specifically aimed at assessing (1) possible asymmetries in the funnel plot as a potential indicator of publication bias, (2) if there was sufficient heterogeneity in the sample to justify further analysis, and (3) if effect sizes from different methods could be combined into a single analysis, to boost power. Analyses were performed with the meta (Schwarzer, 2012) and metafor (Viechtbauer, 2010) packages for R (R Core Team, 2012).

We analyzed funnel plot asymmetry as a potential indicator of publication bias (Egger, Smith, Schneider, & Minder, 1997). In a funnel plot effect sizes are plotted against some measure of study size, and in a symmetric plot large studies are expected to cluster in the middle, while smaller studies are spread to both sides. Figure 1 shows an underrepresentation of studies in the lower left corner, that is, studies with a high

Including these measures in the key regression for this study was not possible, as it would have imposed a serious curfew on our statistical power.

standard error and small effect size. This could occur for a variety of reasons, including that such studies may be set aside before or after the submission stage on the grounds that the sample size is too small. Please note that the rightmost three datapoints are outliers over 3 SD from the sample mean and were excluded from subsequent analyses. A linear regression on funnel plot asymmetry reaches significance [$t(95) = 4.86$, $p < .001$], suggesting bias (publication or otherwise) in our sample. To assess whether the found asymmetry reflected different effect size distributions across methods rather than an overall bias, analyses of funnel plot asymmetry were also conducted separately by method. We found significant asymmetry for all methods except for CHT, with the sample of EEG studies being too small to assess asymmetry. These results are not reported here but available on request.

----Insert Figure 1 about here-----

Figure 1 furthermore gives an indication that experiments cluster by method. We followed up on this observation by assessing the sample characteristics, first overall and then by method. As a first step, we estimated the overall effect size. We chose a random effects model for the analysis, which allows heterogeneity between studies due to differences in, for instance, sample characteristics or method chosen. The mean weighted effect size under a random effects model was $estimate = 0.401$ ($SE = 0.040$), with the lower bound of the 95% confidence interval $CI_L = 0.329$, and the higher bound $CI_H = 0.484$. This effect size was significantly different from zero ($z = 10.25$, $p < .001$). As a second step, we assessed heterogeneity of the sample. Next to estimating the mean true effect, the amount of heterogeneity among the true effects needs to be estimated in a random-effects model. τ^2 measures between-study variance as an estimate of the

difference between total observed variance and within-study variance. The total amount of between-study variance was $\tau^2 = 0.054$ (estimated by restricted maximum likelihood, REML). Expressed in percentages, the variability explained by heterogeneity rather than sampling error was $I^2 = 40.64\%$ [$CI_L = 23.49\%$, $CI_H = 61.14\%$]. Cochran's Q-test for homogeneity indicated significant sample heterogeneity [$Q(96) = 163.426$, $p < .001$]. This result indicates that the sample variance is larger than would be expected from sample error, which justifies the introduction of moderator variables into the analysis.

In order to estimate the variance explained by the experimental method, we conducted a second analysis on overall sample characteristics, introducing experimental method as a moderator variable. The CF method was used as the reference level for this factor, because it has the largest amount of observations (40) and the lowest mean effect size. The Q-test showed significant heterogeneity between methods [$Q(4) = 19.523$, $p < .001$], and the effect of CHT was significant (*estimate* = 0.541, $z = 4.27$, $p < .001$) with a significantly higher mean effect size than CF. Residual heterogeneity remained significant [$I^2 = 81.62\%$, $\tau^2 = 0.035$, $Q(92) = 135.405$, $p = 0.002$], indicating that method did not account for all the variance.

The above analyses show considerable heterogeneity between methods, cautioning us to be careful in combining effect sizes from different experimental methods into one analysis. Moreover, residual heterogeneity also remains considerable, suggesting that the sample contains variability beyond the portion accounted for by method. We therefore included method as a moderator variable. It should also be noted that data on native contrasts ($k = 75$) outnumber data on non-native ones ($k = 22$), as evident in Figure 2.

----Insert Figure 2 about here-----

3.2 Does effect size vary developmentally as a function of whether the contrast is present in the infants' native language?

We entered vowel nativeness (native, non-native), age (in days), and their interaction into the analysis. Given the heterogeneity of effect sizes across methods, method was entered as an additional factor. There is no reason to predict that the relationship between age and nativeness will interact with method; moreover, there are too few points to reliably estimate the slope of the change in native and non-native discrimination as a function of age separately for each method. Therefore, no interactions with method were declared. The categorical factors nativeness and method were contrast-coded. Thus, the intercept estimates the weighted mean effect size at age = 0. The comparison level for method was again CF.

The Q test for moderators was significant [$Q(7) = 32.061$, $p < .001$], showing that the regressors that we included accounted for a substantial proportion of variance. The Q test on residual heterogeneity was also significant [$Q(89) = 119.837$, $p = 0.016$], which indicates that further factors may be needed to account for the remaining variance. The model intercept was significant (*estimate* = .476, *SE* = .112, $z = 4.235$, $p < .001$), suggesting that baseline discrimination levels were significantly different from zero. Additionally, there was a significant interaction between nativeness and age (*estimate* = -.0021, *SE* = 0.0009, $z = -2.356$, $p = .019$), which is consistent with the hypothesis that developmental trends for native and non-native contrasts diverge. The CHT method (*estimate* = .601, *SE* = 0.135, $z = 4.468$, $p < 0.001$) and the HPP method (*estimate* = .

1781, *SE* = 0.089, *z* = 1.986, *p* < 0.047) showed a significant effect. We carried out a number of follow-up analyses to make sure that these results were robust. For the sake of simplicity, we do not report them in detail here. In one set of follow-ups, we assessed the possibility that method accounted for the results found above. To this end, we separated CHT, HPP and other methods, as well as removed the NIRS results; the same pattern of results found in the general analyses obtained in all three regressions. Additionally, we conducted two analyses declaring either study or sound contrast instead of method as a structuring variable. These also replicated the previous results, as the interaction between nativeness and age remained significant in both of them.

3.3 How does discrimination of native contrasts change with age?

We followed up on the divergence in developmental trends by fitting separate models for native and non-native contrasts. For the native contrasts ($k = 75$), the Q test for moderators reached significance [$Q(5) = 19.410$, $p = .002$], suggesting that our regressors were capturing meaningful variation. Additionally, the Q test for residual heterogeneity was also significant [$Q(69) = 91.694$, $p = .035$], indicating that a substantial proportion of variance remained to be explained. In this statistical analysis, the baseline discrimination level again differed from zero, because the intercept reached significance (*estimate* = .364, *SE* = .104). The linear slope for age also reached significance (*estimate* = .001, *SE* = .0004, *z* = 2.249, *p* = 0.025). Additionally, the methods CHT (*estimate* = .593, *SE* = .159, *z* = 3.733, *p* < 0.001), HPP (*estimate* = .250, *SE* = .096, *z* = 2.617, *p* = 0.009), and NIRS (*estimate* = .316, *SE* = .160, *z* = 1.981, *p* = 0.048) showed significant

effects. We conducted additional analyses to assess if age was better captured with quadratic or cubic trends, but neither of these predictors (derived from a centered version of age) had a significant slope in subsequent polinomial regressions.

3.4 How does discrimination of non-native contrasts change with age?

For the non-native contrasts (k = 22), the test for moderators was significant [Q(5) = 17.149, p = .004], whereas the test for residual heterogeneity was not [Q(16) = 16.682, p = .286], suggesting that our regressors succeeded in structuring the variance in the dependent variable. The baseline level of discrimination for non-native contrasts was above zero, as the intercept was significant (*estimate* = .534, *SE* = .196; z = 2.725, p = .006). The slope for CHT was also a significant predictor (*estimate* = .596, *SE* = .234, z = 2.543, p = 0.011), again indicating that effect sizes with this method are substantially higher. The slope for age did not achieve significance, although the estimate was in the predicted negative direction (*estimate* = -.0012, *SE* = .0008, z = -1.439, p = .150). Quadratic and polinomial regressors based on age did not have a significant estimate in this analysis either.

3.5 At what age does vowel perception become language-specific?

Given the interest that there has been for the age of the emergence for language-specific perception, we sought to provide some rough estimation that could be further investigated in future research. There are several possible ways of approaching the question of the age at which attunement occurs. One is to identify the crossover, given a

linear fit was accurate for at least native perception. The crossover of weighted linear regression lines for native and nonnative effect sizes was at 165 days (5.4 months; cf. Fig. 2). Another possibility is to group effect sizes as a function of the age at which the data had been collected. We divided age-groups into 4 quartiles and carried out weighted regressions exactly as those above (declaring nativeness and method) in each of those quartiles, to assess at which age group native and non-native effect sizes diverged. Nativeness did not have a significant estimate in the first two quartiles (3 to 131 days, and 132 to 185 days) but it was a significant predictor of effect size in the later two quartiles, namely between 6 months and 10 months (*estimate* = -.576, *SE* = .285; z = -2.023, p = 0.043), and 10 and 14 months (*estimate* = -.594, *SE* = .268; z = -2.217, p = 0.027).

4. Discussion

In standard theoretical views (including NLM and PRIMIR), discrimination improves for native vowels within the first year of life, whereas it declines for non-native vowels during that time. We carried out a meta-analysis of developmental infant vowel discrimination literature to assess these predictions. Detailed statistical analyses provided evidence for perceptual narrowing in vowels, in the form of an interaction between vowel nativeness and age. This interaction was due to significantly different slopes for native and non-native sounds. Moreover, effect sizes for native vowel discrimination increased significantly with age. Statistically significant evidence for non-native vowel discrimination was not found, a point to which we return below. As for the age at which attunement occurs, significant differences between effect sizes elicited using native and non-native contrasts were apparent in data collected after, but not much before, 5.4 or 6 months of age.

The first conclusion to be drawn from these data is that there is clear statistical support in current developmental vowel discrimination data, from a variety of paradigms, that perception of native and non-native vowels comes to diverge over the first year of life. This conclusion is not trivial in view of the fact that several null results have been reported for changes in perception with age (and thus language exposure and/or across two language backgrounds; e.g., Polka & Bohn, 1996; Sebastián-Gallés & Bosch, 2009). We believe that our results put both positive and negative previous results in a new, holistic perspective of infant perception, as follows.

To begin with, the presence of an interaction between age and nativeness together with an effect of nativeness in datapoints gathered after 6 months confirm the predictions from perceptual attunement in general, and the description made from the NLM and PRIMIR

models in particular. Indeed, enhancement in discrimination of native contrasts had mainly been documented in consonants (Kuhl, Stevens, Hayashi, Deguchi, Kiritani, & Iverson, 2006; Narayan, Werker, & Beddor, 2009; see also Pons et al., 2012), and thus it is compelling that the present meta-analysis, profiting from the power of studies testing over a thousand infants, was able to confirm that the extrapolation of this process to vowels was justified. At the same time, the lack of a significant slope for non-native datapoints taken separately cautions as to both the strength of the effect and the design that should be adopted in the future.

This is especially true because the decline in discrimination of non-native has, in a way, been a stronger tenet in the literature on perceptual narrowing in speech sound contrasts. Early findings of a decline in non-native speech perception (Werker & Tees, 1984) led researchers to assume a universal listener who is able to discriminate all speech sound contrasts in the world, and whose ability to do so declines with language exposure. Only recently have reports of improvement began to appear (Kuhl et al., 2006), resulting in the presently predominant view of both decline and enhancement based on language exposure. Our results suggest that the changes in non-native discrimination are rather small in size, as they cannot be distinguished from the null hypothesis independently.

One possibility we considered related to PAM (Best, 1995), a model discussed briefly in the introduction. In it, non-native contrasts are not all difficult to discriminate. On the contrary, those non-native contrasts that can be mapped onto native ones may remain quite discriminable. For instance, both English and German contrast the vowels [i-ɪ], as in the English words 'sheep' and 'ship'. Although these vowels are not exactly the same across the two languages, the German contrast is quite easy to discriminate by native

American English listeners because the German [i] maps onto their native English [i], and the German [ɪ] maps onto the English [ɪ]. Thus, one may wonder if some of the non-native results might have been of this 'easy' type. Deciding on this would require a relatively extensive study of the infants' native language and the stimuli used, which could be explored in future research. Nonetheless, we are not confident that this analysis is promising, given that the statistic for remaining variance to be explained was not significant. Instead, we suggest that the current null result for the change with age among non-native effect sizes could be due to insufficient power, because we benefited from only 22 non-native compared to 75 native effect sizes. Therefore, future work including non-native contrasts would be desirable to make the native and non-native samples more comparable.

We propose to take these results as indication that a stronger measure of language attunement would be obtained as the difference between *two* discrimination indices from the same children, one for a native contrast and the other for a non-native one. Such a design has already been successfully employed in the study of consonant attunement (Conboy et al., 2005), where investigators cleverly selected a single standard sound as background (voiceless unaspirated /t/) and measured reactiveness to two oddballs. One of the oddbals was contrastive in the infants' native language (either voiced /d/ for Spanish learners, or aspirated /t^h/, for English learners). Such an oddball paradigm is compatible with both CHT and ERPs. This design would also keep a better handle on random acoustic differences across the contrasts tested; that is, to some extent, one could have feared that nativeness effects might have been obscured if all the native sounds employed happened to be more acoustically dissimilar than non-native contrasts. By testing three

sounds in a single continuum or matching the two pairs in acoustic distance, future research would be better able to measure language-specific effects.

Another interesting finding obtained in the present meta-analysis relates to the discussion of whether vowel perception attunes earlier than consonants (e.g., Pons et al., 2012). Our analyses show that perception indeed differs as a function of nativeness as early as 6 to 9 months of age, but not much before this point. We would like to, however, withhold judgment as to whether this age range is earlier for vowels than consonants until the appropriate meta-analysis has been done with consonantal data.

It should be noted that, albeit significant, the effects observed for age are rather small. An analysis on consonantal data would shed light on whether these small attunement effects reflect a minor role of language exposure in shaping perception or rather are peculiar to vowels. As mentioned in the introduction, infants' vowel perception is less pliable in laboratory learning experiments than similar approaches in consonants.

Before concluding, it is relevant to discuss the limitations of the current study. The first three are inherent to meta-analyses, which are only as good as the data they are based on. Thus, one important limitation related to sample size for analyzing the effect of potential modulating factors. Indeed, we could not conduct separate analyses within methods, or even include further moderator variables like acoustic distance between stimuli, acoustic distance of non-native stimuli from native categories, as well as further experimental and stimulus characteristics in a quantitative way.

The second, which must also temper our enthusiasm for the attunement effects described above, relates to the possibility that our data reflects a publication bias which

is, itself, shaped by theoretical expectations. Notice in particular that the great majority of results came from published studies, with only 4 being manuscripts at this point. In our searches, we have not come across theses or reports in conferences, which are more likely to contain null results that are usually not accepted in peer-reviewed journals. As any other meta-analysis, this one is only as truthful as the data it includes. In fact, we found statistical evidence for a bias in our data suggestive that small effect sizes were being under-reported. It should be clarified, however, that this is not akin to a publication bias regarding age and nativeness interactions. That is, our sample is biased towards reporting positive *discrimination* results beyond age and the native/non-native status. Nonetheless, bias remains an important consideration that should be kept in mind, particularly given that only developmental studies (i.e., reporting more than one age group) were included.

A third limitation of the present work relates to the 'apples and oranges' problem constitutive of meta-analysis. This type of research necessarily builds on diverse studies, and ours is no exception. We included here a host of different studies, with variable designs, and which load to a variable extent on discrimination skills per se. For example, CHT studies require of the infant not only that she hears the difference between two tokens, but also that she refrains from making a response when no change has occurred, which undoubtedly involves executive abilities beyond linguistic discrimination. Infants tested in CHT also go through a long period of shaping and are highly trained in the task, whereas infants in, for example, NIRS studies will typically simply be presented with either one or two vowels, with no specific training to perform a discrimination task. This difference could possibly lead to a higher likelihood of finding mixed results, and might

be one reason why effect sizes derived from CHT were significantly higher than those derived from other methods.

A related limitation goes beyond the meta-analytic nature of the present research, and relates to the underlying phenomenon under study. Discrimination has been used as an early index of language acquisition, but the precise mechanisms by which this occur remain poorly understood, as evidenced by the differences across the NLM and PRIMIR models of attunement. Primarily due to limitations in the available data, the current meta-analysis has not taken into account factors such as acoustic distance between vowels or acoustic variability induced by number of tokens or talkers, which are certainly relevant for a more differentiated picture of perceptual narrowing. More in general, we cannot speak to the fundamental question of at what level reorganization occurs. There is considerable evidence from adult studies that we retain sensitivity to non-native contrasts (particularly vocalic ones, e.g., Beddor & Strange, 1982). Such findings have led to the hypothesis that language acquisition operates in a 'structure-building' process, and that cross-linguistic differences in perception are driven by top-down influences, for example through biases induced by certain types of tasks (Schouten, Gerrits, & van Hessen, 2003), whereas lower levels of perception remain completely faithful to the signal (but see Chandrasekaran, Krishnan, & Gandour, 2007 for evidence that language experience can shape even the brainstem's response to non-linguistic sounds). Furthermore, attunement in discrimination is clearly only the first of many steps in the road to the native language. Put into a lexical context, infants do not simply discriminate phonemes along the relevant dimensions to make lexical distinctions, but also attend to indexical information like talker identity (e.g., Houston & Jusczyk, 2003; Rost & McMurray, 2010). Even within

speech perception alone, infants must also gain a host of other abilities and considerable knowledge at many other levels of representation (e.g., Werker, Fennell, Corcoran, & Stager, 2002, Fernald, Perfors, & Marchman, 2006). These interesting questions go well beyond the present meta-analysis, although they may be amiable to future ones in which more automatic (i.e., EEG, NIRS) and more "decision-based" (i.e., CHT) discrimination responses can be directly compared.

To conclude, we sought experimental evidence concerning the emergence of native language perception patterns for vowels in infancy. A meta-analysis supported the contention that native and non-native discrimination develop in opposite directions over the first year of life. Moreover, a distinction is evident already by about 6 months of age. In addition to substantiating claims made from mainstream models (NLM and PRIMIR), the present results suggested that a fruitful future avenue of research could employ multiple measures for better capturing infants' budding linguistic knowledge.

Notes

We are grateful to Laura Bosch, Yasuyo Minagawa, Ferran Pons, Yutaka Sato who provided us with further information on their published studies; to Titia Benders, Liquan Liu, René Kager, and Reiko Mazuka for making their unpublished manuscripts available to us; to Olusola Adesope, Kimmo Alho, Carl Dunst, and Frans van der Slik for helpful discussion regarding the meta-analytic methods; to Minna Huotilainen for helpful discussion of studies to consider; and to Amanda Seidl and Derek Houston for unpublished data included in earlier versions of this manuscript. This work has also benefited greatly from discussions with other colleagues at MPI, LSCP, RU, and the Dutch Baby Circle. All remaining errors are our own.

References

(References marked with * were included in the meta-analysis)

Beddor, P. S. & Strange, W. (1982). Cross-language study of perception of the oral-nasal distinction. Journal of the Acoustic Society of America, 71, 1551-1561.

Benders, T. (submitted). Learning phonemes from multiple auditory cues: Dutch infants' language input and perception. *

Best C.T., McRoberts G.W., LaFleur R., & Silver-Isenstadt J. (1995). Divergent developmental patterns for infants' perception of two nonnative consonant contrasts. Infant Behavior and Development, 18(3), 339-350. doi: 10.1016/0163-6383(95)90022-5

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In J. C. Goodman & H. C. Nusbaum (Eds.), The Development of Speech Perception: The Transition from Speech Sounds to Spoken Words (pp. 167–224). Cambridge, MA: The MIT Press.

Bonatti, L. L., Peña, M., Nespor, M., & Mehler. J. (2004). Linguistic constraints on statistical computations: The role of consonants and vowels in continuous speech processing. Psychological Science, 16, 451-459.

Bosch, L., & Sebastián-Gallés, N. (2003). Simultaneous bilingualism and the perception of a language-specific vowel contrast in the first year of life. Language and Speech, 46, 217–243. doi: 10.1177/00238309030460020801 *

Caramazza, A., Chialant, D., Capasso, D., & Miceli, G. (2000). Separable processing of consonants and vowels. Nature, 403, 428-430.

Chandrasekaran, B., Krishnan, A., & Gandour, J. T. (2007). Experience-dependent neural plasticity is sensitive to shape of pitch contours. *Neuroreport, 18*, 1963-1967.

Cheour, M., Alho, K., Ceponiené, R., Reinikainen, K., Sainio, K., Pohjavuori, M.,... Näätanen, R. (1998). Maturation of mismatch negativity in infants. International Journal of Psychophysiology, 29(2), 217–26. *

Cheour, M, Leppänen, P.H. T., & Kraus, N. (2000). Mismatch negativity (MMN) as a tool for investigating auditory discrimination and sensory memory in infants and children. Clinical Neurophysiology, 111(1), 4-16. doi: 10.1016/S1388-2457(99)00191-1

Cristia, A., McGuire, G. L., Seidl, A., & Francis, A. L. (2011). Effects of the distribution of acoustic cues on infants' perception of sibilants. Journal of Phonetics, 39(3), 388-402. doi: 10.1016/j.wocn.2011.02.004

Cristia, A., Seidl, A, Junge, C., Soderstrom, M., & Hagoort, P. (in press). Predicting individual variation in language from infant speech perception measures. Child Development.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. BMJ, 315(7109), 629–634. doi:10.1136/bmj.315.7109.629

Feldman, N.H., Griffiths, T.L., & Morgan, J.L. (2009). The influence of categories on perception: Explaining the Perceptual Magnet Effect as optimal statistical inference. Psychological Review, 116(4), 752-782. doi: 10.1037/a0017196

Fernald, A., Perfors, A., & Marchman, V. A. (2006). Picking up speed in understanding. Developmental Psychology, 42, 98-116.

Figueras Montiu, M., & Bosch Galceran, L. (2010). Capacidades de discriminación fonética de un contraste vocálico nativo en el prematuro. Psicothema, 22(4), 669-676. *

Frieda, E. M., Walley, A. C., Flege, J. E., & Sloane, M. E. (1999). Adults' perception of native and nonnative vowels: Implications for the perceptual magnet effect. Attention, Perception, & Psychophysics, 61(3), 561-577. doi: 10.3758/BF03211973

Granier-Deferre, C., Ribeiro, A., Jacquet, A.-Y., Bassereau, S. (2011). Near-term fetuses process temporal features of speech. Developmental Science, 14(2), 336-352. doi: 10.1111/j.1467-7687.2010.00978.x

Grieser, D., & Kuhl, P. K. (1989). Categorization of speech by infants: Support for speech-sound prototypes. Developmental Psychology, 25(4), 577–88. doi: 10.1037/0012-1649.25.4.577

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. The Journal of the Acoustical society of America, 97(5), 3099-3111. doi: 10.1121/1.409456

Houston, D. M., & Jusczyk, P. W. (2003). Infants long-term memory for the sound patterns of words and voices. Journal of Experimental Psychology: Human Perception and Performance, 29(6), 1143-1154. doi: 10.1037/0096-1523.29.6.1143

Jansson-Verkasalo, E., Ruusuvirta, T., Huotilainen, M., Alku, P., Kushnerenko, E., et al. (2010). Atypical perceptual narrowing in prematurely born infants is associated with compromised language acquisition at 2 years of age. BMC Neuroscience, 11(1), 88. doi:10.1186/1471-2202-11-88 *

Jusczyk, P. W. (1993). From general to language-specific capacities: The WRAPSA model of how speech perception develops. Journal of Phonetics, 21(1-2), 3–28.

Kemler Nelson D. G., Jusczyk P. W., Mandel D. R., Myers J., Turk A., & Gerken, L. (1995). The head-turn preference procedure for testing auditory perception. Infant Behavior and Development, 18(1), 111-116. doi 10.1016/0163-6383(95)90012-8

Kuhl, P. K. (1991). Human adults and human infants show a perceptual magnet effect for the prototypes of speech categories, monkeys do not. Perception & psychophysics, 50(2), 93–107. doi: 10.3758/BF03212211

Kuhl, P. K. (1994). Learning and representation in speech and language. Current Opinion in Neurobiology, 4(6), 812–822. doi:10.1016/0959-4388(94)90128-7

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). Philosophical transactions of the

Royal Society of London. Series B, Biological sciences, 363(1493), 979–1000. doi:10.1098/rstb.2007.2154

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. Developmental Science, 9(2), F13-F21. doi: 10.1111/j.1467-7687.2006.00468.x

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. Science, 31(255), 606–608. doi: 10.1126/science.1736364

Lipsey, M. W. & Wilson, D. B. (2001). Practical meta-analysis. Thousand Oaks, CA: SAGE.

Liu, L., & Kager, R. (2011). How do statistical learning and perceptual reorganization alter Dutch infant's perception to lexical tones? ICPhS XVII, 1270-1273. Retrieved from http://www.icphs2011.hk/resources/OnlineProceedings/RegularSession/Liu,%20Liquan/Liu,%20Liquan.pdf

Liu & Kager (in preparation a). Infants' perceptual development towards a native vowel contrast. *

Liu & Kager (in preparation b). Bilingual infants' perceptual development towards a native vowel contrast. *

Marean, G., Werner, L., & Kuhl, P. K. (1992). Vowel categorization by very young infants. Developmental Psychology, 28(3), 396-405. doi: 10.1037/0012-1649.28.3.396 *

Mazuka, Hasegawa, & Tsuji (submitted). Development of non-native vowel discrimination: Improvement without exposure. *

Minagawa-Kawai, Y., Mori, K., Naoi, N., & Kojima, S. (2007). Neural attunement processes in infants during the acquisition of a language-specific phonemic contrast. The Journal of neuroscience, 27(2), 315–21. doi:10.1523/JNEUROSCI.1984-06.2007 *

Minagawa-Kawai, Y., Naoi, N., Nishijima, N., Kojima, S., Dupoux, E. (2007). Developmental Changes in Cerebral Responses to Native and Non-Native Vowels: A NIRS Study. Proceedings of the International Conference of Phonetic Sciences XVI. *

Mugitani, R., Pons, F., Fais, L., Dietrich, C., Werker, J. F., & Amano, S. (2009). Perception of vowel length by Japanese- and English-learning infants. Developmental psychology, 45(1), 236–47. doi: 10.1037/a0014043 *

Narayan, C., Werker, J. F., & Beddor, P. S. (2009). The interaction between acoustic salience and language experience in developmental speech perception: evidence from nasal place discrimination. Developmental Science, 13(3), 407-420. doi: 10.1111/j.1467-7687.2009.00898.x

Polka, L, & Bohn, O.-S. (1996). A cross-language comparison of vowel perception in English-learning and German-learning infants. The Journal of the Acoustical Society of America, 100(1), 577-592. doi: 10.1121/1.415884

Polka, L., & Bohn, O.-S. (2011). Natural Referent Vowel (NRV) framework: An emerging view of early phonetic development. Journal of Phonetics, 39(4), 467–478. doi: 10.1016/j.wocn.2010.08.007 *

Polka, L., & Werker, J. (1994). Developmental changes in perception of nonnative vowel contrasts. Journal of experimental psychology. Human perception and performance, 20(2), 421–435. doi: 10.1037/0096-1523.20.2.421

Pons, F., Mugitani, R., Amano, S., & Werker, J. F. (2006). Distributional learning in vowel length distinctions by 6-month-old English infants. Presented at the International Conference on Infant Studies; Kyoto, Japan (abstract).

Pons, F., Sabourin, L., Cady, J. C., & Werker, J. F. (2006). Distributional learning in vowel distinctions by 8-month-old English infant. Presented at the 28th Annual Conference of the Cognitive Science Society; Vancouver, BC, Canada (abstract).

Pons, F., Albareda-Castellot, B., & Sebastián-Gallés, N. (2012). The interplay between input and initial biases: asymmetries in vowel perception during the first year of life. Child Development, 83(3), 965–76. doi: 10.1111/j.1467-8624.2012.01740.x *

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rost, G. C., & McMurray, B. (2010). Finding the signal by adding noise: The role of non-contrastive phonetic variability in early word learning. Infancy, 15(6), 608-635. doi: 10.1111/j.1532-7078.2010.00033.x

Sato, Y., Sogabe, Y., & Mazuka, R. (2010). Discrimination of phonemic vowel length by Japanese infants. Developmental Psychology, 46(1), 106–119. doi: 10.1037/a0016718 *

Sato, Y., Mori, K., Furuya, I., Hayashi, R., Minagawa-Kawai, Y., & Koizumi, T. (2003). Developmental changes in cerebral lateralization during speech processing measured by near infrared spectroscopy. Japanese Journal of Logopedic Phoniatrics, 44, 165–171. doi: 10.5112/jjlp.44.165. *

Schouten, B., Gerrits, E., & van Hessen, Arjan (2003). The end of categorical perception as we know it. Speech Communication, 41, 71-80.

Schwarzer, G. (2012). meta: Meta-Analysis with R. R package version 2.1-3. Retrieved from http://CRAN.R-project.org/package=meta

Tsuji, S., & Cristia, A. (in preparation). InPhonDB: A developing meta-analysis of infant vowel perception.

Sebastián-Gallés, N., & Bosch, L. (2009). Developmental shift in the discrimination of vowel contrasts in bilingual infants: is the distributional account all there is to it? Developmental Science, 12(6), 874–87. doi: 10.1111/j.1467-7687.2009.00829.x *

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. Journal of Statistical Software, 36(3), 1-48. URL http://www.jstatsoft.org/v36/i03/

Weikum, W. M., Oberlander, T. F., Hensch, T. K., & Werker, J. F. (2012). Prenatal exposure to antidepressants and depressed maternal mood alter trajectory of infant speech perception. PNAS, 109, 17221-17227. doi: 10.1073/pnas.1121263109

Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. Language Learning and Development, 1(2), 197-234. doi: 10.1080/15475441.2005.9684216

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant behavior and development, 7*(1), 49-63. doi: 10.1016/S0163-6383(84)80022-3

Werker, J. F., Cohen, L. B., Lloyd, V. L., Casasola, M., & Stager, C. L. (1998). Acquisition of word-object associations by 14-month-old infants. *Developmental Psychology, 34*(6), 1289–309.

Werker, J. F., Fennell, C. T., Corcoran, K. M., & Stager, C. L. (2002). Infants' ability to learn phonetically similar words. *Infancy 3(1),* 1-30.

Werker, J. F., Polka, L., & Pegg, J. E. (1997). The conditioned head turn procedure as a method for testing infant speech perception. *Early Development and Parenting, 6*(3-4), 171–178. doi:10.1002/(SICI)1099-0917(199709/12)6:3/4<171::AID-EDP156>3.0.CO;2-H

Figure captions

Figure 1: Funnel plot of effect sizes by method. Different methods are represented with different symbols, as shown in the legend.

Figure 2: Effect size as a function of age, nativeness, and method. Different methods as well as nativeness are represented with different colors and symbols, as shown in the legend. Lines indicate meta-analytic regression of effect size by age fitted to the relevant set of points. These lines do not take method into account.