How does the human auditory system become expert in speech processing? Insights from development.

Cécile Issard and Alejandrina Cristia

31 07, 2019

- Target journal: Developmental science
- Article type: short report
- 4000 words
- 6 keywords
- Running title: 40 characters
- Submit one normal and one blinded version
- Separate files for title page, main text, and figures
- No identifiying info in the main text.
- up to 4 research highlights; each 25 words
- Abstract: 250 words

Main text file:

1. Title
2. Research highlights
3. Abstract and key words
4. Main
5. References
6. Figures and tables (each clearly identified, labelled and on a separate page)
7. Appendices (if relevant).

Abstract

The human auditory system is amazingly efficient at processing speech. This capacity would be present from birth, infants preferring to listen to natural speech than to other types of sounds, enabling them to select the signals that are relevant for communication with homospecifics. However, a large variety of sounds have been contrasted to speech, with infants of very different ages. Drawing a global picture of how this capacity emerges is therefore difficult. We synthesized the literature by conducting a meta-analysis of studies testing speech preference in infants from birth to one year of age. We found a strong effect size, infants prefering speech over any other type of sound. However, contrary to the results of individual studies, we found no effect of age: infants showed the same amount of preference from birth to one year of age. Preference was stronger when speech was contrasted to artificial sounds, and when the speech stimuli were in the infants' native language. This suggests that the representation of speech as a distinct auditory object emerges from a broader category of natural sounds, modulated by the degree of familiarity with the sound.

Introduction

Speech is probably the most important sound class for humans. It is the main signal for vocal communication, and as such it is crucial that individuals detect this sound in the environment to spot a homospecific and build social interactions. Readily from birth, humans would be equipped with a capacity to recognize speech sounds, to process them with dedicated auditory and cognitive mechanisms. At birth, infants discriminate speech from complex tones (Dehaene-Lambertz, 2000), and sine-wave speech (SWS) (Vouloumanos & Werker, 2007). Extending to language acquisition, the naturalness of sound stimuli (i.e. using synthetic vs. natural speech) is a key factor for infants to segment words (Black & Bergman, 2016). This highlights the importance of recognizing natural speech sounds in the auditory environment to trigger the relevant cognitive processes for this stimulus, including for language acquisition. Discriminating speech from other sounds, and preferring it over other types of sound, may be a necessary condition to learn language. Here we synthesize empirical data on infants' preferences for speech over artificial sounds, natural sounds, as well as human and social non-speech sounds.

A key question is whether speech is preferred per se, or because it belongs to a broader category of natural or own-species sounds. Studies from the auditory neuroscience literature have provided evidence that natural sounds are processed preferentially by the auditory system, from the cochlea (Lewicki, 2006) to the auditory cortex (e.g. Gehr et al., 2000) (see Mizrahi et al., 2014 for a review). Consistently, infants have been shown to discriminate between speech and various types of artificial sounds from birth, from white-noise (Colombo, 1981) to sine-wave speech (Vouloumanos et al., 2007), low-pass filtered speech (Cooper, 1994), and backward speech (Peña, 2003; May et al, 2011, 2018). This preference is maintained to the end of the first year of life (Curtin, 2013; Vouloumanos, 2014). The infant auditory system would thus detect general acoustical properties that differentiate artificial from natural sounds, among them speech. However, studies contrasting speech to other natural sounds have nuanced this view. Newborns made more head-turns to speech than to heartbeat (Ecklund-Flores & Turkewitz, 1996), but listened equivalently to speech and monkey calls (Vouloumanos & Werker, 2010). Infants younger than 3 months have been shown to not discriminate between speech and monkey calls, whereas infants from 3 months of age do (Vouloumanos et al., 2004). It is thus possible that infants rely on a more specific category for vocal sounds, that includes our closest genealogical cousins (i.e. primates), whose vocalizations may share some important acoustical properties with speech. In an fMRI study, the activity of the temporal cortex in response to biological non-speech sounds (such as human non-speech vocalizations or rhesus calls) decreased with age between 1 and 4 months-old. The response to speech didn't increase during the same developmental window (Shultz et al. 2014), suggesting that the capacity to discriminate speech from other sounds comes from a narrowing of the perceptual category to speech rather than a more in-depth processing. Infants therefore appear to discriminate vocal from other natural sounds from the beginning of their life, and later to discriminate and preferentially process speech specifically as they get older.

But is it really an effect of age, or does preference come from familiarity with specific sounds that infants frequently encounter in their environment? Larger hemodynamic responses were observed in the newborn brain for forward as compared to backward speech when the native language was used for the speech stimuli, but not when a foreign language was used (Sato et al., 2012; May et al., 2018). In 4 month-old infants, speech produced similar activation patterns for speech and non-speech vocal sounds, with a larger difference when the speech stimuli were in the native language of the participants (as compared to when speech was in a foreign language) (Minagawa-Kawai et al., 2011). This suggests that infants use their knowledge of the language they are familiar to to discriminate speech from other sounds. Furthermore, 9 month-old infants listen longer to monkey calls than their native language (Sorcinelli, 2019), possibly because at this age they are already attuned to the sounds of their native language and transfer their attention to the more demanding sound in the paradigm.

Finally, it is possible that the infants' auditory system preferentially process vocal sounds, and later speech, because they share a complex acoustical structure. Indeed, studies comparing speech to music often found a lack of preferential processing. At two months, the temporal cortex showed the same amount of repetition suppression for speech and music (Dehaene-Lambertz et al., 2010). Interestingly, this lack of discrimination persists even after infants discriminate speech from other complex vocal sounds: at five months, infants detected speech or music equivalently well in an auditory scene (anonymous, 2019). It is therefore possible that the developing auditory system is attuned to complex specific parameters shared by music and speech, but not in other animal vocalizations.

The complex developmental pattern that we describe above is even more difficult to understand that controversies exist in the literature. When speech was compared to other human sounds, two different laboratories found different patterns: In the first case, speech triggered larger BOLD responses and looking times than other human sounds, both communicative (e.g. laugh or agreement) and non-communicative (e.g. yawns or coughs) in 1 to 4 month-old infants (Shultz et al., 2014, Shultz & Vouloumanos, 2010). In the second case, the opposite pattern was observed: human communicative and non-communicative sounds evoked larger responses than speech in a similar region at 3 months. No difference between the three types of sounds was observed in the same infants at 6 months (MacDonald et al., 2019). Moreover, even studies that found consistent results contrasted speech to a large variety of sounds, from white noise to filtered speech, and at different ages. Getting a precise overview of this capacity is therefore difficult. This points to the importance of synthesizing the available data is a systematic way.

All of these results have been observed on small groups of infants, with a large variety of age and stimuli across groups. Individual studies can only test a few infants on very specific stimuli due to experimental constraints (Oakes, 2017; Bergmann et al., 2017; Sugden & Moulson, 2015). On the opposite, meta-analysis are a way of achieving power without running new studies. They gather data from significantly more infants than individual studies, which significantly increases statistical power. By merging a lot of different studies, they allow researcher to state support (or not) for some results with controversy, and provide statistical evidence of how much we can trust the results. If an effect emerges, it's more likely to be a reproducible one that different labs can find. Finally, meta-analysis offer tools to detect publication bias in the literature, providing even more evidence to support or not the results (i.e. significant effects being less trustworthy if they emerge from a biased literature). However, meta-analysis have the disadvantage of mixing studies with different experimental designs together, therefore having less control on the effect measured. They merge results focusing on common factors between studies, potentially missing subtle effects. But moderators analyses allow to explain the heterogeneity between study. In individual studies, addressing questions such as differences across stimuli and age groups types would require large power. Meta-analysis allow to draw a developmental timeline across the age range covered by the literature, and test how the different factors discussed by different individual studies interact. For all of these reasons, we conducted a meta-analysis to test if infants' reliably have a preference for speech sounds over other types of sounds, and if yes, if different types of sound modulated this preference, and how it developed over the first year of life.

Methods

Literature search

We followed PRISMA (Moher et al., 2009). The information sources used to compose the initial list included suggestions by experts (authors of this work); two google scholar searches (“ (“speech preference” OR “own-species vocalization”) AND infant”, and “(”speech preference" OR “own-species vocalization”) AND infant”) complemented with the same searches in PubMed and PsycInfo; and a google alert, as well as reference lists of the full papers inspected. After a first screening based on titles and abstracts, we ran a second round of screening based on full paper reading.

Inclusion criteria

We included studies that tested human infants from birth to 1 year (0-365 days) of age, and contrasted speech sounds with any other type of sound, measuring either behavioral (e.g. looking times) or neurophysiological responses to the sounds. We excluded studies that contrasted foreign to native language, didn't present natural speech sounds, presented speech recorded with the mother's voice, or intentionally mixed speech with other vocal sounds within the same sound condition. We included published (e.i. journal articles) as well as unpublished works (e.i. doctoral dissertations). A PRISMA flow chart summarizes the literature review and selection process (Figure 1). We documented all the studies that we inspected in a decision spreadsheet (supplementary materials).

[Insert Figure 1 here]

Data were coded by the first author. 20% of the papers were selected to be coded by the second author independently, with disagreements resolved by discussion. There were **XX** disagreements out of a total of **YY** fields filled in, so that the total agreement rate was **ZZ**%. For effect sizes, we coded the mean score and the standard deviation for each sound condition. When individual data was provided, we recomputed the respective mean scores and standard deviations based on the reported individual scores. When they were reported, we coded the t-statistic between the two sound conditions or the F-statistic. If a Cohen's d oran Hedge's g effect size was directly reported we also coded this. The full list of the variables coded is available in the supplementary material.

Risk assessment at the level of papers was done by ... Risk assessment for the whole body of literature ...

Statistical analysis

Individual effect sizes

Once the data were coded, we computed individual effect sizes that were not directly reported in the papers, along with their respective variance. We adjusted the formula according to the experimental design of the respective paper (Lipsey & Wilson, 2001). When the coded study used a within-participant design with two measurements (e.g. looking time during speech and during monkey calls), we computed effect size using t-statistic (Dunlap et al., 1996). If this statistic was not reported, we computed effect size based on the respective means and SDs. We then corrected the computed effect size with the correlation between the two measurements. We computed this correlation based on the t-statistic, the respective means and SDs (Lipsey & Wilson, 2001). If not all of these informations were reported, we randomly imputed a correlation with equal probability between 0.01 and 0.99.

Effect sizes were first computed as Cohen's d, and then transformed to Hedge's g.

Meta-analytic models

Once the data was completed, we estimated the true effect size fitting mixed-effects meta-analytic regressions. We used the R package metafor **(CITE)**. We specified a hierarchical model with random effects of paper, and random effects for independent infants within paper (same_infant). We specified the following moderators as fixed effects: - mean age of children; - experimental method (Central fixation/Head-turn Preference Procedure/High Amplitude Sucking/Passive Listening); - familiarity with the language used (native/foreign); - naturalness of the contrastive sound (natural/artificial, coded as yes/no). If the sound contrasted to speech was natural, we also coded whether it was vocal or not, and from human or another species (homospecific/heterospecific, coded as yes/no).

We first estimated the global effect size by fitting a random-effects meta-analytic regression without any hierarchical structure or moderator. We then added the hierarchical structure with papers and infant groups within papers. We assessed whether the experimental method influenced the magnitude of the effect size apart from target moderators by fitting a mixed-effects meta-analytic regression with the above described hierarchical structure and the method as a moderator.

We investigated the effect of familiarity with the sound by running a mixed-effects meta-analytic model with nativeness of the language used for the speech stimuli as a moderator.

We then investigated whether speech preference was embedded in a preference for natural sounds, and whether this potential effect evolved over the course of the first year of life, by fitting a mixed-effects meta-analytic model with naturalness and age as moderators. To facilitate result interpretation, we centered age.

Finally, we subsetted the dataset to contrasts between speech and natural sounds, and fitted a meta-analytic regression on this subset with vocalness (vocal/non-vocal), and species (homospecific/heterospecific) as moderators.

Publication bias

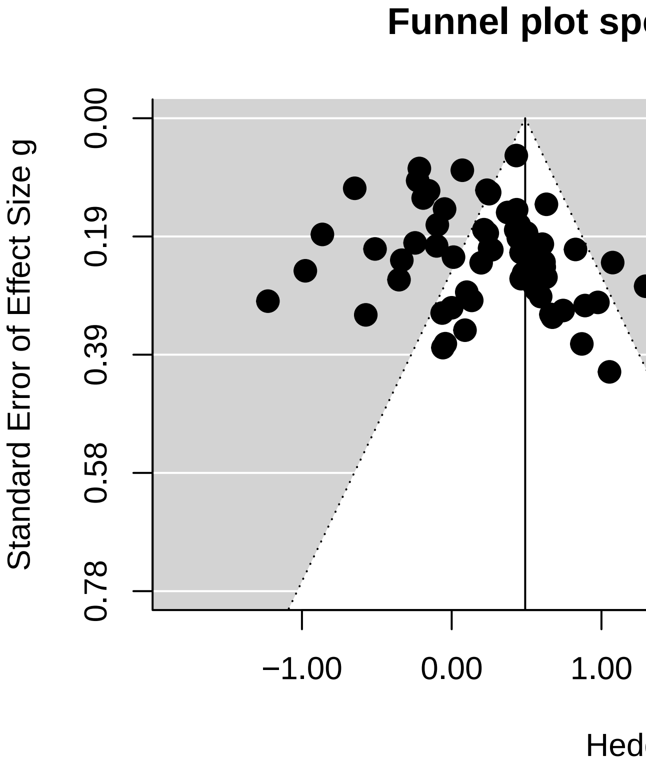
We assessed the presence of a potential publication bias in the literature by plotting N funnel plots. The first one was based on the simple model without hierarchical structure nor moderators. We symmetrized this funnel plot using the "trim and fill" method (ADD REF). The second funnel plot was based on the mixed model with hierarchical random structure and moderators. We tested the asymmetry of the funnel plots by regressing effect size as a function of effect size standard error and running a Kendall's tau rank test.

Results

Database description

We found a total of 25 papers reporting 92 (not mutually independent) effect sizes, see Table 1. All of them have been submitted to or published in peer-reviewed journals. Studies tended to have small sample sizes, with a median N of 20.5 children (Range = 53, M = 21.4565217, Total: 874. Infants ranged from 0 to 12 (1.46 to 380.5 days). Individual samples comprised 29 % of female participants on average. Infants were native of 7 different languages across the whole database (English, French, Japanese, Italian, Russian, Yiddish, Hebrew). Studies were performed in 13 different laboratories from 6 different countries (United States, Canada, Israel, France, Japan, Italy). 4 experimental methods were used: 12 studies used Central Fixation (CF), 3 used High-Amplitude Sucking (HAS), 1 used Head-turn Preference Procedure (HPP), and 9 used Passive Listening (PL).

Publication bias



Evidence of bias at level of papers Evidence of bias at level of literature

```
## pdf
##   2
```

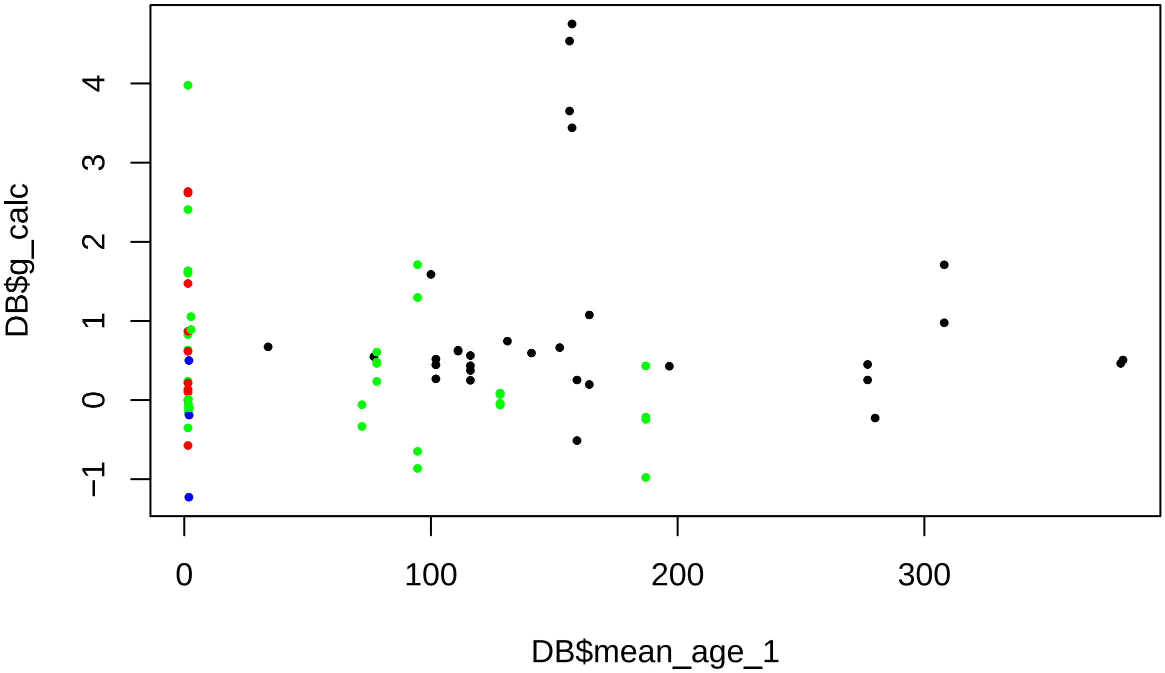
```
##
## Regression Test for Funnel Plot Asymmetry
##
## model:     mixed-effects meta-regression model
## predictor: standard error
##
## test for funnel plot asymmetry: z = 9.0361, p < .0001
```

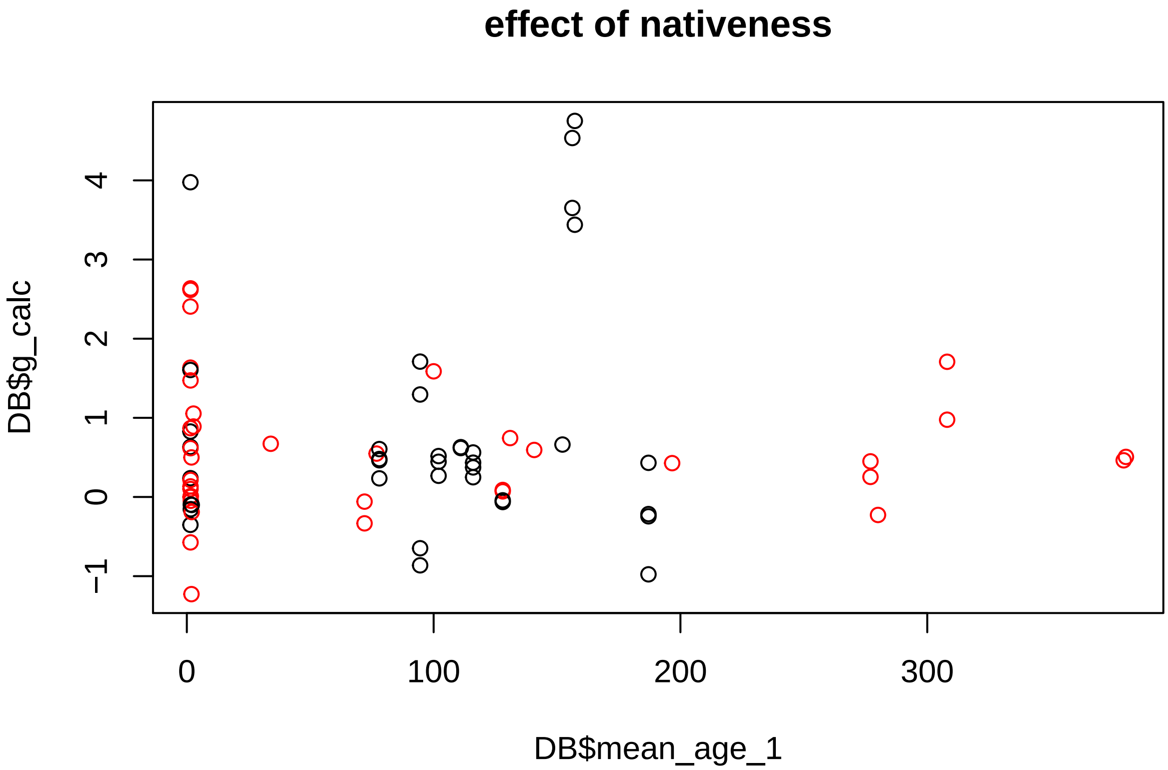
```
##
## Rank Correlation Test for Funnel Plot Asymmetry
##
## Kendall's tau = 0.3438, p < .0001
```

Main effects

Including Plots

```
##      CF      HAS      HPP      PL
## "black"  "blue"   "red" "green"
```

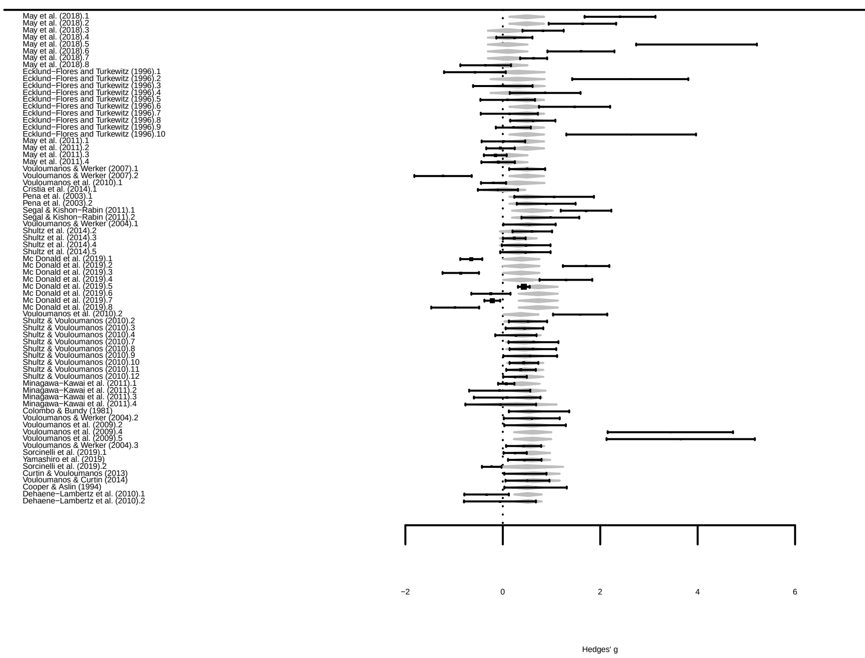




```
##   foreign    native
## 0.8075473 0.5997470

##   foreign    native
## 1.4358134 0.8595385

##
##           no yes
##   foreign  9  26
##   native  28   8

##
## Multivariate Meta-Analysis Model (k = 71; method: REML)
##
##    logLik   Deviance        AIC        BIC       AICc
## -188.5627   377.1255   397.1255   418.5568   401.3563
##
## Variance Components:
##
##             estim    sqrt  nlvls  fixed                factor
## sigma^2.1  0.1261  0.3552     21     no              study_ID
## sigma^2.2  0.1347  0.3670     35     no  study_ID/same_infant
##
## Test for Residual Heterogeneity:
## QE(df = 63) = 576.2598, p-val < .0001
##
## Test of Moderators (coefficient(s) 2:8):
## QM(df = 7) = 23.0223, p-val = 0.0017
##
## Model Results:
##
```

```
##                            estimate      se     zval    pval    ci.lb
## intrcpt                      0.3606  0.1919   1.8790  0.0602  -0.0155
## natural1                    -0.1247  0.3324  -0.3752  0.7075  -0.7763
## test_lang2                   0.0050  0.2328   0.0215  0.9828  -0.4512
## agec                         0.0038  0.0010   3.6547  0.0003   0.0017
## natural1:test_lang2          0.2836  0.3974   0.7138  0.4754  -0.4952
## natural1:agec               -0.0022  0.0037  -0.6031  0.5464  -0.0094
## test_lang2:agec             -0.0031  0.0022  -1.3793  0.1678  -0.0074
## natural1:test_lang2:agec     0.0019  0.0042   0.4538  0.6500  -0.0063
##                             ci.ub
## intrcpt                    0.7366    .
## natural1                   0.5268
## test_lang2                 0.4612
## agec                       0.0058  ***
## natural1:test_lang2        1.0625
## natural1:agec              0.0050
## test_lang2:agec            0.0013
## natural1:test_lang2:agec   0.0101
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Multivariate Meta-Analysis Model (k = 30; method: REML)
##
##    logLik   Deviance        AIC        BIC       AICc
## -118.1860   236.3720   248.3720   255.9206   252.7931
##
## Variance Components:
##
##             estim    sqrt  nlvls  fixed               factor
## sigma^2.1  0.2655  0.5153     10     no             study_ID
## sigma^2.2  0.5964  0.7723     15     no  study_ID/same_infant
##
## Test for Residual Heterogeneity:
## QE(df = 26) = 355.3470, p-val < .0001
##
## Test of Moderators (coefficient(s) 2:4):
## QM(df = 3) = 12.6078, p-val = 0.0056
##
## Model Results:
##
##                     estimate      se     zval    pval    ci.lb   ci.ub
## intrcpt               0.5714  0.3095   1.8460  0.0649  -0.0353  1.1781  .
## homospecific2        -0.1088  0.1853  -0.5873  0.5570  -0.4720  0.2543
## agec                  0.0032  0.0041   0.7757  0.4379  -0.0049  0.0112
## homospecific2:agec    0.0004  0.0042   0.0875  0.9302  -0.0079  0.0086
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Forest plot of effect sizes

We found a mean weighted effect size g=0.78 (CI[,]) Heterogeneity Moderators age, type & interaction

Discussion

- Age
- Type of sounds contrasted
- Interactions
- power
- heterogeneity

Naturalness alone doesn't significantly moderate infants' preference for speech sounds: they still prefer speech, and the amount of preference doesn't change, whether the sound is natural or artificial. This means that infants prefer natural speech in itself. This preference might explain why naturalness makes a difference for higher-level linguistic, a priori abstract tasks such as word segmentation (Black & Bergmann, 2016): speech triggers different cognitive mechanisms than other sounds. Not incompatible as in this meta-analysis natural speech stimuli when contrasted to only synthetic speech. We don't know if infants would have been able to find "words" with natural sounds other than syllables (i.e. frequent sequences of several natural sounds).

Experimental method significantly modulates speech preference: some methods are more appropriate than other to test infants on this type of task. This phenomenon has been repeatedly observed in developmental meta-analysis (see Bergmann et al., 2018, for a synthesis across meta-analysis in developmental psychology).