Which acoustic and phonological factors shape infants' vowel discrimination? Exploiting natural variation in InPhonDB

Sho Tsuji[1]*, Alejandrina Cristia*[2]

[1]University of Pennsylvania, Philadelphia, USA
[2]LSCP, Département détudes cognitives, ENS, EHESS, CNRS, PSL Research University, Paris, France.

tsujish@gmail.com, alejandrina.cristia@ens.fr

Abstract

A key research question in early language acquisition concerns the development of infants' ability to discriminate sounds, and the factors structuring discrimination abilities. Vowel discrimination, in particular, has been studied using a range of tasks, experimental paradigms, and stimuli over the past 40 years, work recently compiled in a meta-analysis. We use this meta-analysis to assess whether there is statistical evidence for the following factors affecting effect sizes across studies: (1) the order in which the two vowel stimuli are presented; and (2) the distance between the vowels, measured acoustically in terms of spectral and quantity differences. The magnitude of effect sizes analysis revealed order effects consistent with the Natural Referent Vowels framework, with greater effect sizes when the second vowel was more peripheral than the first. Additionally, we find that spectral acoustic distinctiveness is a consistent predictor of studies' effect sizes, while temporal distinctiveness did not predict effect size magnitude. None of these factors interacted significantly with age. We discuss implications of these results for language acquisition, and more generally developmental psychology, research.

Index Terms: speech recognition, language acquisition, meta-analyses, vowel discrimination, phonology, phonetics, asymmetries

1. Introduction

In 1971, Eimas and colleagues [1] reported that 1-month-old human infants responded differently to an acoustic change spanning an adult-perceived phonemic boundary (between English-like /d/ and /t/) than to a change with the same acoustic distance not spanning this boundary. This astounding finding spurred a wealth of experimental research seeking to document the initial or experience-independent sensitivities infants display as well as the incredible strides they make towards native-like speech perception within the first years of life. Much experimental work used vocalic stimuli, not only to study early language acquisition (e.g., [2, 3]), but also to describe the development of the auditory and/or general cognitive system (e.g., [4, 5]).

All of this experimental work assessing infants' discrimination of vocalic sounds begs the question: What factors explain structured variance in infants' discrimination of these sounds? Taking up the example of Eimas' work, together with other research, it is now apparent that there are certain discontinuities in the perception of acoustic space, likely as a side effect of Mammalians' auditory systems, which are exploited by languages by placing consonantal category boundaries in certain regions that make the contrast easily distinguishable not only to Eimas' infants but also to adult humans who are given a more-fine grained choice [6] and to chinchillas even without extensive supervised training [7]. However, such discontinuities are not apparent in adult perception of vocalic sounds. To our knowledge there is relatively little research directly addressing this question (with one exception to be discussed next). There is, however, an obvious reason why researchers may be unable or unmotivated to carry out this work: To assess the effect of acoustic dissimilarity, it becomes necessary to test multiple contrasts. Given infants' limited attention spans, this typically means additional infant groups, one per contrast tested – thus placing a very hard requirement on researchers. And yet, one would certainly like to make sure that effect sizes from infant research reflect obvious patterns, such that, for instance, greater acoustic distances (all else equal) lead to greater discriminability effects.

The one exception to the general pattern whereby each paper reports results on a single sound contrast is studies documenting order effects, also called asymmetries, which have been argued to follow from discontinuities in perception in a theoretical framework called the Natural Referent Vowel framework [8]. NRV states that more peripheral vowels are better perception anchors than less peripheral ones. As a result, change detection in the direction from less peripheral towards more peripheral vowels is easier and stronger than vice versa. Effects consistent with this description have been found in a range of experimental studies (qualitative reviews in [9, 8]); for instance, Pons and colleagues [10] studied the discrimination of [i-e] contrast by Spanish- and Catalan-learning infants. Both groups of infants were better at discriminating the contrast in the direction from [e] to [i] at 4 and 6 months of age. However, other findings pattern in unexpected ways. For instance, Mazuka and colleagues [11] document asymmetries in directions opposite to those predicted by NRV for [i-e] among Japanese 10-month-olds and [o-u] among 4-month-olds.

To sum up, individual studies are insufficient to answer, in an objective and quantitative way, the key questions of whether vowel discrimination displays reliable order effects, and to what extent discriminability is predicted by acoustic distances between the vowel stimuli used. However, these questions could be answered by a meta-analytic approach, which compiles the statistical power of a whole research field. Indeed, it is possible to extract effect sizes from all papers assessing discrimination of vocalic sounds. Effect sizes can be expressed in standardized metrics, such as Cohen's d, which is a measure of signal to noise potentially allowing for cross-paper comparisons. This technique of extracting comparable effect sizes across a body of work is called meta-analysis, and it may be described as a tool to bring together diverse studies in a broader analysis conceptually encompassing them all.

In previous work [12], we have compiled a database con-

taining all public studies testing discrimination of vocalic sounds by infants. Here, we assess whether there is statistical evidence in this public database for the following factors affecting effect sizes across studies: (1) the order in which the two vowel stimuli are presented; and (2) the distance between the vowels, measured acoustically in terms of spectral and quantity differences. It is conceivable that the effects of these variables are modulated by native language experience. Therefore, we systematically include interactions with infant age and whether the contrast is native (i.e., present in the ambient language) or not.

2. Methods

We drew from the InPhonDB meta-analysis, available from MetaLab `metalab.stanford.edu`. Since the construction of the meta-analysis has been described in general terms elsewhere [12], we provide here only some general statistics about the database today, given that it has been updated after publication [13]. At present, InPhonDB contains data from 2735 infants (mean age = 233 days, range 3-912 days) from 39 papers, collectively containing 191 effect sizes.

Whether one expects a greater effect given NRV was coded on the basis of the vowels' position in F1/F2 space. If the first vowel presented was less peripheral than the second vowel, this was coded as "yes" (because it should be easier to discriminate when the order is less to more peripheral), or "no" if the opposite order was used. Studies in which experimenters counterbalanced order of presentation and did not report discrimination results for each order separately are not considered for the asymmetry analyses.

Additionally, where reported, each experiment was coded in terms of the acoustic characteristics of the stimuli used, particularly the two vowels' positions in F1/F2 space as well as their duration. From this information, a spectral distinctiveness was calculated by first transforming F1/F2 into the bark scale and then taking the square root of the sum of the squared distances found in the F1 and F2 dimensions; and the temporal distinctiveness as the ratio between the shortest and the longest vowel.

We excluded points where nativeness was ambiguous (N = 6), either because it was an allophonic distinction or because the contrast was instantiated in a dialect that was different to the infants' own. We further excluded experiments where the sound stimuli were paired with a visual object to concentrate on sound discrimination, since these studies potentially tap word-object association learning (N = 5). We also excluded experiments where infants were not typically developing and monolingual (N = 29), and data points whose effect sizes were more than 3 standard deviations from the mean of the meta-analysis (N = 4), since subsequent analyses could be affected by such outliers. Since some of these exclusion criteria overlapped, the final dataset contained N = 152 data points. This dataset was restricted further depending on the analysis, as noted below.

Based on our prior work with this database, we know it is necessary to control for whether studies involved a habituation phase, a fixed-length familiarization phase, a conditioning phase, or neither, since effect sizes are markedly different as a function of this methodological aspect. Thus, in addition to infant age and nativeness of the contrast, we included method as a predictor in the base model. We used a hierarchical random effects model accounting for the fact that data points stemming from the same paper might share more variance than data points stemming from different papers. Within each paper, we added random effects for each data point stemming from an independent infant group.

Our dependent measure is a standardized effect size corrected for small sample sizes, Hedges' g. As a test of statistical significance, we compared base models to full models assessing goodness of fit, as indicated by likelihood ratio tests. We used multivariate meta-analytic regression models using the metafor package [14] in R [15]. Our base models had the structure *model = effect size, effect size variance , mods = nativeness * age + exposure phase, random = ~1 |paper/infant, weighted = TRUE*; and the full models added the effect of the respective predictor to this base model. In addition, we inspected the model summaries of the respective full models for interactions of the target predictor with age or nativeness.

Analysis scripts and data can be retrieved on the project osf site at `https://osf.io/px885/`

3. Results

3.1. Peripherality

We were able to include 26 data points where discrimination was tested from a less peripheral to a more peripheral vowel, and 19 data points for the opposite directionality. Model comparison showed that there was a significant effect of peripherality (see Table 1; Fig.1). We additionally explored the estimated mean effect sizes for the two types of data points by constructing separate models. The intercept for the model with data points testing discrimination from less to more peripheral vowels was higher [b = 0.902, $p \leqslant .001$, CI_L = 0.674, CI_L = 1.130] than the intercept for the model with data points testing discrimination in the opposite direction [b = 0.475, $p \leqslant .001$, CI_L = 0.287, CI_L = 0.664]. Meta-analytic results thus support the predictions of the NRV model. We do, however, want to caution that the dataset is relatively small, due to the fact that not many studies report discrimination results split by direction. Since it is possible that studies would only report directional effects in case their data support the NRV model, we assessed funnel plot asymmetry as an indicator of publication bias [16]. We based this funnel plot on difference scores for studies that report discrimination results bidirectionally, with each score being the difference between the effect sizes for the respective directions. Egger's regression test did not reach significance [z = -0.327, p = .744]; see also Fig. 1. Inspection of full model summaries revealed a marginally significant interaction of age and peripherality [b = -0.002, p = .053, CI_L = -0.004, CI_L = 0.000]. However, follow-up models suggest no age effect on discrimination from less to more peripheral vowels [b = 0.000, p = .991, CI_L = -0.016, CI_L = 0.017], or in the opposite direction [b = 0.001, p = .244, CI_L = -0.001, CI_L = 0.003].

3.2. Spectral and temporal distance

A total of 107 data points had associated information on **spectral distance**. Adding this predictor significantly increased model fit (see Table 1, Spectral (all); Fig. 2).There were no interactions with nativeness or age.

Temporal distance could be obtained for 90 data points. This predictor did not improve model fit, (Table 1, Temporal (all); Fig. 2), and there were no interactions with nativeness or age.

We carried out several additional analyses to make sure that our results were not due to biases in data selection or unfair comparisons. To verify that differences in outcomes between our analyses of spectral and temporal distance were not based

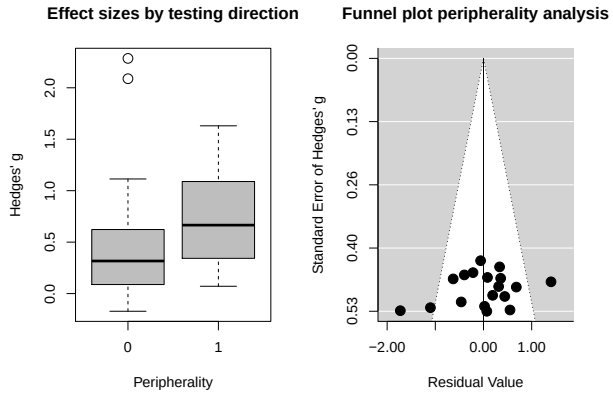

Figure 1: *Left side: Boxplot of Hedges' g effect sizes by peripherality status. "No" includes contrasts tested from more to less peripheral, and "Yes" includes contrasts tested from less to more peripheral. Right side: Funnel plot of model residuals against standard errors of effect sizes as a potential indicator of publication bias towards NRV-compatible peripherality findings. The vertical line indicates the median effect size estimate based on the model, and the white area indicates a pseudo confidence interval region around this value with bounds equal to ± 1.96 SE.*

on differences in datasets (107 vs. 90 data points), we conducted subset analyses based only on the 62 data points containing information on both measures. Both analyses on spectral and temporal differences yielded results consistent with the main analysis (Table 1, sub-overlap results). Finally, although we statistically account for the fact that data points from the same paper could be more similar than data points from different papers, it might still be the case that comparisons drawn from the same paper provide a more powerful measure, as presumably they control for differences in e.g., laboratory habits. In a second subset analysis, we included papers that provided multiple data points for a given dimension, for instance two vowel contrasts. The spectral subset analysis contained 77 data points, and the temporal analysis contained 53 data points. Again, we see a significant improvement in model fit for spectral, but not temporal distance (Table 1, sub-multiple results).

4. Discussion

The present study sought to assess whether general descriptors of vowel contrasts predicted infant performance in the current body of experimental literature. Our first conclusion bears on the general enterprise, as the fact that we do find significant meta-analytic regressors suggests that infant experimental data, collapsed across dozens of studies, remains sensitive enough to structuring factors. In other words, when one employs such a big data approach, one may be gaining power by aggregating across the results of literally hundreds of infants, but also increasing noise by comparing "apples" from one laboratory against "oranges" in another laboratory. The current pattern of results suggests that the former may make up for the latter, at least for certain factors that have been sufficiently studied in the literature and/or sufficiently strong in their main effects. We will acknowledge certain limitations of this approach, however, in the detailed discussions pertaining our two specific research

Table 1: *Results of model comparisons for analyses on peripherality, spectral, and temporal distance. LRT = Likelihood ratio test. Sub-overlap indicates that the analysis has been run on studies where information on both spectral and temporal characteristics was provided. Sub-multiple indicates that analysis has been carried out on studies reporting two or more levels of the relevant dimension (e.g., in spectral, where multiple spectral contrasts were tested within the same paper).*

Predictor	LRT	df	*p*-val
Peripherality (all)	11.813	4	**.019***
Spectral (all)	19.236	4	**>.001***
Spectral (sub-overlap)	13.338	4	**.010***
Spectral (sub-multiple)	13.828	4	**.008***
Temporal (all)	3.677	4	.452
Temporal (sub-overlap)	3.337	4	0.503
Temporal (sub-multiple)	7.162	4	0.128

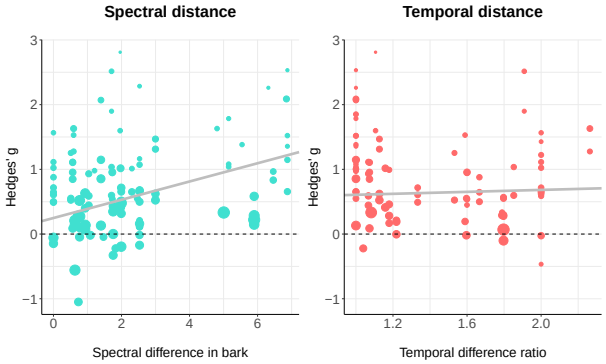

Figure 2: *Effect size as a function of spectral distances (left panel) or temporal distances (right panel). Each point corresponds to one experiment, with size representing the weight of that experiment in the meta-analytic regression. Data portrayed corresponds to the full data for each analysis.*

sub-questions, which follow.

4.1. Peripherality effects

We confirm one general prediction of NRV, finding greater effect sizes when sound discrimination is tested from the less to the more peripheral vowel, compared to the opposite order, with no interaction with nativeness, and only a marginally significant interaction with age that is not backed by follow-up analyses. Regarding the main peripherality effect, some readers may wonder whether it might not simply reflect selective reporting, e.g. that researchers are more likely to report orders separately when they conform to NRV predictions than when they do not. Moreover, a non-negligible amount of the data points included in this analysis (14/45 data points, or 18% of regression weight) are authored by the originators of the NRV model. However, we do not think this is a likely explanation, since our test on funnel plot asymmetry does not support selective reporting of NRV-compatible effects. It may still be the case that we overestimate the size of the effect, for instance if authors tend not to report order effects at all when they do not find a significant difference.

The main effect of peripherality is fascinating for theories of infant perception and language acquisition, as it sug-

gests that infants' perceptual system is sensitive to these order effects regardless of experience. Peripherality effects are also interesting for modeling work, given that they provide a cognitive validation parameter. To our knowledge, few papers proposing or evaluating a model for the acquisition of sound categories explicitly address language-independent order effects. For instance, Feldman and colleagues [17] attempt to integrate all asymmetries within the general framework of experience-induced Native Language Magnet effects (as defined by e.g. [2]). Such proposals are likely unable to explain why asymmetric discrimination is evident extremely early on, having been repeatedly documented at 2-4 months [3, 18] and thus well before the onset of natively-attuned vowel discrimination (meta-analytically established at 6 months by [12] using a subset of the same global data set employed here).

We add here that the NRV framework as most recently specified [8] predicts precisely such asymmetries early on, but additionally proposes that language experience overrides these effects, with adults and older infants displaying stronger peripherality effects for non-native than native contrasts. Our results do not show the expected triple interaction (peripherality x nativeness x age). This may well be because such interaction is beyond our statistical power; in the smaller cell for the two-way interaction peripherality x nativeness there are only 9 data points. We hope that further experimental work may boost our meta-analytic power, to assess the strength of the empirical evidence behind NRV's developmental predictions.

4.2. Spectral and temporal distance

Our analyses overall do not find evidence for an effect of temporal distance on effect sizes, whereas they confirm that spectral distance predicts effect sizes significantly.

Our results for the simple spectral distance measure we employed are encouraging, and suggest there is sufficient variability in the dataset to explore the matter further. We implemented distance as the Euclidean separation on F1-F2 space in Bark, but many other instantiations are possible. It would be extremely interesting to assess the predictive value of more linguistically informed representations (see e.g., [19]), but unfortunately few authors make their stimuli available in their raw form, thus limiting their re-description. Another direction of work we have not explored pertains to the relative importance of different formants, which has been proposed in one word-learning study, potentially in a post-hoc manner, whereby F1 would be perceptually more salient than F2 [20]. Our data are available from `metalab.stanford.edu`, and thus interested readers can download them and assess for themselves this and hopefully many other hypotheses.

In contrast to the robust effects of spectral distance, temporal distances were not a significant predictor in any of the analyses. Might this indicate a true difference between the dimensions, or a chance finding due to confounding factors? To better evaluate this marked difference in outcomes, let us consider to what extent differences in the spectral and temporal variability accessible in our database could contribute to these results. Examining the standard deviations of the respective predictor, we find spectral distance to have a larger standard deviation (*sd* = 1.99) than temporal distance (*sd* = 0.40). A second way to look at this question is to examine how well controlled the respective other dimension was in the two sets of data. We found that 19% (20/107) of data points in the spectral analysis contained, in addition to spectral changes, a length difference that could be phonologically contrastive. In contrast, 60% (52/86; note that featural distance could not be coded for 4 data points that had a temporal distinction) of data points in the temporal analysis contained, in addition to a temporal difference, a quality difference that could be phonologically contrastive (such as backness or height). Thus, if anything, the spectral set was more controlled for temporal variation than the temporal set was for spectral variation. Third, we inspected whether changes were greater for spectral versus temporal contrasts in our dataset. Indeed, 90% (96/107) of data points in the spectral analysis actually contained one or more featural changes, while this was the case for only 41% (35/86) of data points in the temporal set, which more often contained within-category variation. Together, these observations suggest that the spectral distance measure had a higher variability, was better controlled for the other dimension, and was more likely to test a phonologically-contrastive featural difference than the temporal distance measure.

Despite these empirical concerns, we also note one conceptual difference between the two dimensions, namely that spectral distance covers a wider spectrum of phenomena in natural language. That is, length is easily captured in a unidimensional acoustic or phonological feature, while encoding spectrum differences requires several phonological features and/or more complex, multidimensional physical representations (such as 2 or more formants, or multiple mel bands). Within a given dimension, moreover, vowel variation in natural languages seems to span larger distances for spectral than temporal contrasts. Drawing from our own dataset, the maximal spectral difference is provided by the point vowel contrast [a-i], spanning 6 bark; whereas the maximum temporal difference found here corresponds to a ratio of duration 2 (i.e., double the length), corresponding to a lexically contrastive length contrast in Japanese. Together with the various subset analyses we have carried out, these conceptual considerations lead us to conclude that the difference in meta-analytic effects measured here for spectral versus temporal differences may indeed be due to actual differences in perceptual effects across these two dimensions, although further work with more carefully controlled data could conclude otherwise.

5. Conclusions

Using a meta-analytic approach, we confirmed piece-wise reports that order of presentation impacts discriminability, leading to recommendations for experimentalists to systematically report effect sizes separating counterbalanced order, and for modelers to attempt to assess whether their models can accommodate such order effects prior to (and perhaps despite) native language experience. We also found that vocalic contrasts spanning a greater spectral distance led to greater effect sizes than those spanning smaller distances, whereas the same could not be said for contrasts varying in duration. We hope these interesting results will motivate experimentalists to share the raw stimuli used in their experiments, which will allow more fine-grained analyses.

6. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the under the Marie Sklodowska-Curie grant agreement No 659553, and the Agence Nationale pour la Recherche [ANR-14-CE30-0003 MechELex, ANR-10-IDEX-0001-02 PSL*, ANR-10-LABX-0087 IEC].

7. References

[1] P. D. Eimas, E. R. Siqueland, P. Juscyk, and J. Vigorito, "Speech perception in infants," *Science*, vol. 171, no. 3968, pp. 303–306, 1971.

[2] P. K. Kuhl, K. A. Williams, F. Lacerda, K. N. Stevens, and B. Lindblom, "Linguistic experience alters phonetic perception in infants by 6 months of age." *Science*, vol. 255, no. 5044, pp. 606–8, 1992.

[3] L. Polka and J. F. Werker, "Developmental changes in perception of nonnative vowel contrasts." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 20, no. 2, p. 421, 1994.

[4] N. Cowan and P. A. Morse, "The use of auditory and phonetic memory in vowel discrimination," *The Journal of the Acoustical Society of America*, vol. 79, no. 2, pp. 500–507, 1986.

[5] L. J. Trainor and R. N. Desjardins, "Pitch characteristics of infant-directed speech affect infants ability to discriminate vowels," *Psychonomic Bulletin & Review*, vol. 9, no. 2, pp. 335–340, 2002.

[6] D. B. Pisoni, R. N. Aslin, A. J. Perey, and B. L. Hennessy, "Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 8, no. 2, p. 297, 1982.

[7] P. K. Kuhl and J. D. Miller, "Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli," *The Journal of the Acoustical Society of America*, vol. 63, no. 3, pp. 905–917, 1978.

[8] "Natural Referent Vowel (NRV) framework: An emerging view of early phonetic development, author=Polka, Linda and Bohn, Ocke-Schwen, journal=Journal of Phonetics, volume=39, number=4, pages=467–478, year=2011, publisher=Elsevier."

[9] L. Polka and O.-S. Bohn, "Asymmetries in vowel perception," *Speech Communication*, vol. 41, no. 1, pp. 221–231, 2003.

[10] F. Pons, B. Albareda-Castellot, and N. Sebastián-Gallés, "The interplay between input and initial biases: Asymmetries in vowel perception during the first year of life," *Child Development*, vol. 83, no. 3, pp. 965–976, 2012.

[11] R. Mazuka, M. Hasegawa, and S. Tsuji, "Development of non-native vowel discrimination: Improvement without exposure," *Developmental Psychobiology*, vol. 56, no. 2, pp. 192–209, 2014.

[12] S. Tsuji and A. Cristia, "Perceptual attunement in vowels: A meta-analysis," *Developmental Psychobiology*, vol. 56, no. 2, pp. 179–191, 2014.

[13] S. Tsuji, C. Bergmann, and A. Cristia, "Community-augmented meta-analyses: Toward cumulative data assessment," *Perspectives on Psychological Science*, vol. 9, no. 6, pp. 661–665, 2014.

[14] W. Viechtbauer, "Conducting meta-analyses in R with the metafor package," *Journal of Statistical Software*, vol. 36, no. 3, pp. 1–48, 2010. [Online]. Available: http://www.jstatsoft.org/v36/i03/

[15] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2016. [Online]. Available: https://www.R-project.org/

[16] M. Egger, G. D. Smith, M. Schneider, and C. Minder, "Bias in meta-analysis detected by a simple, graphical test," *Bmj*, vol. 315, no. 7109, pp. 629–634, 1997.

[17] N. H. Feldman, T. L. Griffiths, and J. L. Morgan, "The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference." *Psychological Review*, vol. 116, no. 4, p. 752, 2009.

[18] K. Wanrooij, P. Boersma, and T. Van Zuijen, "Fast phonetic learning occurs already in 2-to-3-month old infants: An ERP study," *Frontiers in Psychology*, vol. 5, p. 77, 2014.

[19] C. Richter, N. H. Feldman, H. Salgado, and A. Jansen, "A framework for evaluating speech representations," in *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*, 2016.

[20] S. Curtin, C. Fennell, and P. Escudero, "Weighting of vowel cues explains patterns of word–object associative learning," *Developmental Science*, vol. 12, no. 5, pp. 725–731, 2009.