Which acoustic features support the language-cognition link in infancy: A machine-learning approach

Joseph C.Y. Lau (josephcylau@northwestern.edu)
Department of Psychology, Northwestern University,
Evanston, IL 60208, U.S.A.

Alona Fyshe (alona@ualberta.ca)
Department of Computing Science and Psychology, University of Alberta
Edmonton, AB T6G 2R3, Canada

Sandra R. Waxman (s-waxman@northwestern.edu)
Department of Psychology, Northwestern University,
Evanston, IL 60208, U.S.A.

Abstract

From the ambient auditory environment, infants identify which communicative signals are linked to cognition. By 3 to 4 months of age, they have already begun to establish this link: listening to their native language and to non-human primate vocalizations supports infants' core cognitive capacities, including object categorization. This study aims to shed light on the specific acoustic properties in these vocalizations which enable their links to cognition. We constructed a series of supervised machine-learning models to classify those vocalizations that support cognition from those that do not, based on classes of acoustic features derived from a collection of human language and non-human vocalization samples. The models highlight a potential role for *spectral envelope* and *rhythmic* features from both human languages and non-human vocalizations. Results implicate a potential role of underlying perceptual mechanisms relevant to *spectral envelope* and *rhythmic* features in infants' establishment of the uniquely human language-cognition link.

Keywords: Infant cognition, language, non-linguistic vocalizations, acoustic analysis, machine learning

Introduction

Language is fundamental to our species. It offers considerable cognitive power, permitting us to go beyond the here-and-now to establish mental representations of the past and present, and to communicate with others the content of our minds (Miller, 1990). This link between language and cognition is evident even in the first months of infants' life (Perszyk & Waxman, 2018).

For infants as young as 3 to 4 months of age, listening to their native language supports object categorization (Ferry, Hespos, & Waxman, 2010), a core cognitive capacity that is instrumental to learning. Categorization is essential because it permits learners to generalize information about one individual to other members of the same category (Murphy, 2004). This supports rapid learning, permitting learners to extend information efficiently from one individual to new ones, based on their membership within an object category. Categorization also supports memory and reasoning, guiding predictions about novel objects (Gelman, 2004).

Recent evidence reveals that listening to several other acoustic signals with comparable spectral composition (e.g. time-reversed speech), pitch and duration (e.g. sine-wave tone sequences) offer no such cognitive advantage. Moreover, for linguistic stimuli, infants' links to cognition are shaped by their language experience. For 3- to 4-month-old infants acquiring English, listening to either English or German (a "typological cousin" to their native English, with similar prosody) also facilitates object categorization. In contrast, listening to Cantonese (a language from the typologically distant Sino-Tibetan language family) fails to support categorization (Perszyk & Waxman, 2019). Apparently, then, infants' increasingly precise perceptual and neural attunement (i.e., perceptual tuning) to their native language (Kuhl & Rivera-Gaxiola, 2008; Peña, Pittaluga, & Mehler, 2010; Werker, 2018; Werker & Tees, 1984) has powerful downstream consequences beyond perception alone; this perceptual tuning sets boundaries on which other language(s) support infant cognition.

Surprisingly, however, even as infants narrow the range of human languages they link to cognition, vocalizations of non-human primates (e.g., blue-eyed black lemur, *Eulemur macaco flavifrons*) – but not birds (e.g., zebra-finches, *Taeniopygia guttata*) – confer the same cognitive advantage as does listening to their native language (Ferry, Hespos, & Waxman, 2013; Woodruff Carr, Perszyk, & Waxman, 2021). By 6 months, infants have tuned this link; listening to lemur vocalizations no longer supports categorization (Ferry et al., 2013).

How can we best interpret this striking pattern of evidence? Why do 4-month-old infants, who have begun to forge a rather precise link *within* human languages, still maintain an apparently broader link that includes non-human primate vocalizations? In the current paper, we consider these questions by focusing on what acoustic information, available in the surface of the input, might identify a signal as a candidate link to cognition, and on whether the surface properties available to identify candidate links from linguistic vocalizations are the same as, or different from, the properties available in non-human vocalizations.

Implementing a supervised machine-learning (ML) approach, we considered the types of human and non-human vocalizations that, from the vantage point of 4-month-old

English-acquiring infants, either support infant cognition (English, German, lemur vocalizations) or fail to do so (Cantonese, zebra finch vocalizations). Our model was designed to focus on three well-documented classes of acoustic information, including *spectral envelope* features (thought to represent aspects of vocal tract configurations, such as consonant and vowels in speech, e.g. Andén & Mallat, 2014), *rhythmic* features, and *intonational* features (two fundamental elements of speech prosody, e.g. Nooteboom, 1997). Using these three classes of acoustic features, ML classification models were performed to test **a)** whether the models could be trained to make classifications that successfully distinguish vocalizations that support cognition from those that do not, and **b)** which class(es) of acoustic features support that classification.

A total of four sets of classification models were performed (see Figure 1). We first performed classifications using all classes of features combined together in a single inclusive model (*full model*). Performance of the *full model* will identify whether these acoustic properties distinguish vocalizations that support infant cognition from those that do not. We then performed three more specific classifications, each using one of the three features classes (i.e., *spectral envelope*, *rhythmic*, or *intonational* features). Performance of these models will identify which classes of acoustic features, if any, successfully distinguish vocalizations that support cognition from those that do not.

Each classification model was conducted over three sets of vocalization inputs (Table 1). The first input set, which took as its input *human languages*, was trained to classify (English and German) vs. (Cantonese). The second input set, which took as its input *non-human vocalizations*, was trained to classify vocalizations of (lemurs) vs. (zebra finch). The third input set, which took as its input human languages and non-human vocalizations *combined*, was trained to classify (English, German and lemur vocalizations) vs. (Cantonese and zebra finch vocalizations). Models' performance (i.e. classification success) was evaluated by their statistical significance above chance-level, using a permutation approach. Qualitative comparisons in models' performance using these three input sets will offer insight into which classes of acoustic features are instrumental for each input set.

Predictions

Consider first the predictions for the two individual vocalization input sets, i.e. *human languages* and *non-human vocalizations*, each on its own.

We predicted that the *full model* on each of these input sets would both perform successful classifications.

We predicted that the *spectral envelope* models would not perform successful classifications on *human languages*. This prediction is based on evidence that infants younger than 6 to 8 months have not yet narrowed their speech perception sensitivities to focus on segmental (consonant and vowel) contrasts native to their language (Kuhl & Rivera-Gaxiola, 2008; Werker, 2018). In contrast, for *non-human vocalizations*, we expected that the *spectral envelope models* might successfully perform classifications. This prediction is based on the possibility that the link between non-human vocalizations and cognition may be restricted to our closest phylogenetic relatives, whose vocalizations are most similar to human language both articulatorily and perceptually (Perszyk & Waxman, 2018). Articulatory properties, especially vocal tract configurations that differ across species, are richly represented in *spectral envelope* features.

Table 1: Input sets: Human languages (English, German, Cantonese); Non-human vocalizations (Lemur, Finch)

		Vocalizations supporting cognition?	
		Yes	No
Input Sets:	*Human Languages*	English + German	Cantonese
	Non-human vocalizations	Lemur	Finch
	Combined	English + German + Lemur	Cantonese + Finch

We predicted that *rhythmic* models would both perform successful classifications on each of the two input sets. This prediction is based on evidence that speech prosody, which includes rhythm, is instrumental in neonates and young infants' perception of speech (Christophe, Mehler, & Sebastián-Gallés, 2001; Gleitman & Wanner, 1982). This, coupled with evidence of rhythmic features in non-human animal vocalizations (Kotz, Ravignani, & Fitch, 2018), suggests that rhythm may play a role in identifying a candidate link to cognition in both *human language* and *non-human vocalization* input sets, respectively.

We predicted that *intonation* models would also perform successful classifications on each of the two input sets respectively. This prediction is based on evidence that intonation, another component of prosody, contributes to infant perceptual tuning (Chong, Vicenik, & Sundara, 2018). Moreover, there is evidence that intonation-like pitch modulations express emotion and communication intent among in primates (Filippi, 2016). Therefore, we also predicted successful classifications respectively on both *human language* and *non-human vocalization*.

Consider next the success of the models based on the *combined* vocalization input set that takes human languages and non-human vocalizations together as input. Models using the *combined* vocalization input set should illuminate common surface properties, if any, by which candidate links to cognition from linguistic vocalizations and non-human vocalizations, combined, can be identified.

If common surface properties can be identified for the combined vocalization inputs, then the respective models (i.e.

full, *spectral envelope*, *rhythmic*, or *intonational* models) will yield successful classifications.

In contrast, if common surface properties cannot be identified for the combined vocalization input set, then the respective models (i.e. *full*, *spectral envelope*, *rhythmic*, or *intonational* models) will fail to yield successful classifications.

Methods

Materials

Our modeling dataset consisted of a total of 3197 audio samples (Table 2) of human languages and non-human vocalizations for which links to cognition (or the lack thereof) have been attested behaviorally thus far in 4-month-old infants (Ferry et al., 2010, 2013; Perszyk & Waxman, 2019; Woodruff Carr et al., 2021).

Audio samples of human languages were utterance-length recordings produced by multiple female native speakers of each language using an infant directed speech register in interactions with a young child. These included languages in American English (Moser et al., 2020), German (Zahner, Schönhuber, Grijzenhout, & Braun, 2016), and Hong Kong Cantonese (Wang, Kalashnikova, Kager, Regine, & Patrick, 2021). Non-human vocalizations included audio samples of vocalizations of lemurs (Mercer, 2012) and zebra finches (Laboratory of Vocal Learning at Hunter College, 2015). Descriptive statistics of the dataset are presented in Table 2.

Table 2: Input sets: Descriptive statistics of dataset for vocalizations that do (+) and do not (−) support object categorization, from the vantage point of 4-month-old English-acquiring infants

	Vocalization	Label	N=	Duration (sec)
				Mean (SD)
Human	English	+	703	1.23 (0.78)
	German	+	369	2.62 (1.95)
	Cantonese	−	1634	1.94 (0.99)
Non-human	Lemur	+	122	1.55 (0.48)
	Finch	−	369	9.54 (4.59)

Acoustic feature extraction

All audio samples were first normalized in intensity (80 dB). Next, we identified seven acoustic measures that have been shown to primarily represent [1] *spectral envelope*, *rhythmic*, or *intonational* information (e.g. Moser et al., 2020). The full list of measures is presented and described in Figure 1. *Spectral envelope* (MFCC and WTS1), *rhythmic* (ENV, IMF, WTS2, and TMS), and *intonational* (f0) features were created from these seven measures extracted from each vocalization sample.

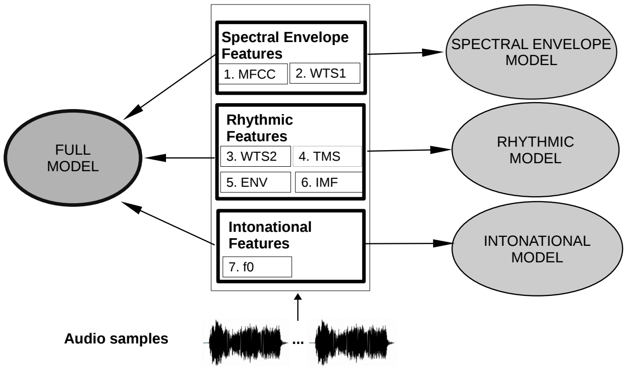

Figure 1: Figure 1: Acoustic measures derived from audio samples: **1.** The mel-frequency cepstral coefficients (**MFCC**) are cepstral representations of the audio sample that concisely describe the overall shape of a spectral envelope as perceived by human. MFCC features were taken as the average of 13 coefficients derived using a 40-band filter bank (133-6864 Hz) on hamming windows of 500ms, each overlapping for 250ms; **2 & 3.** The wavelet time scattering (WTS) transform represents low-variance time-frequency properties that capture the spectral envelope of sounds related to segmental features (**WTS1**), and larger-scale structures like amplitude and frequency modulations relevant to rhythm (**WTS2**). WT1 and WTS2 were taken as averages of log-transformed scattering features, respectively computed using a first wavelet filter bank with 8 wavelets per octave, and a second wavelet filter bank with one wavelet per octave, over non-overlapping windows of 1 sec; **4.** The temporal modulation spectrum (**TMS**) is the frequency decomposition of the temporal envelope of a signal that is a primary acoustic correlate of perceived rhythm in speech, derived using the method described in Ding et al. (2017); **5 & 6.** The speech envelope spectrum (**ENV**) represents temporal regularities correlating to rhythmic properties. Intrinsic mode functions (**IMF**) were decompositions of the ENV using empirical mode decomposition. IMFs represent time-scales of oscillation in the envelope that reflect both syllable-driven and prosodic-feet-level fluctuations. ENV and IMF were derived using the methods described in Tilsen and Arvaniti (2013); **7.** The **f0** contour is a major acoustic correlate of pitch melody and speech intonation, derived from taking 20 f0 values from the raw f0 contour taken at equal proportional intervals. Further, *spectral envelope* (MFCC and WTS1), *rhythmic* (ENV, IMF, WTS2, and TMS), and *intonational* (f0) acoustic measures were input in combination to the Full Model, and individually into respective models corresponding to the three classes of acoustic measures.

[1] We acknowledge that each feature does not only represent information of the class of acoustic feature it is assigned to.

Classification Pipeline

The support vector machine (SVM) classifier was used to perform ML classification. SVM performs binary classification of the data according to pre-specified labels. SVMs perform well even when data is of high dimension and is thus powerful in making classifications in series of multivariate acoustic features which vary in domains such as time, frequency, and amplitude domains. In each classification, a SVM classifier was trained to classify labels corresponding to vocalizations that support cognition and those that do not, based on each audio sample's acoustic feature input. Since the sample size of each type of vocalizations varied (c.f. Table 2), a random undersampling procedure was performed and repeated for 1000 iterations to ensure that the sample size of each category (N=120) was uniform, and that each type of vocalizations under each category was equally represented. A Monte-Carlo cross validation procedure was performed in each of the 1000 iterations, by using 75% of the resample data as the training set, and the other 25% as the testing set. To reduce the dimensionality of the data, a nested feature selection procedure using the Fisher method was then performed on the training set to select 10% of features most informative for the classification. Then, SVM hyperparameter tuning was performed using a Bayesian Optimization procedure, selecting the combination of optimal box constraint, kernel scale, and SVM kernel (Gaussian, linear, or polynomial) which achieved the lowest classification error in a nested 5-fold cross-validation procedure within the training set. An SVM model was then trained on the whole training set using the optimal hyperparameters. Classification performance of the cross validation (based on the model predictions of the testing set labels) of each iteration was quantified as the Area Under the Curve (AUC) of a Receiver Operating Characteristics curve. Statistical significance of each classification was assessed using a permutation approach. A null distribution of AUC values was computed by repeating the same cross-validation procedure for 1000 iterations with the labels of vocalization samples randomized each iteration. The percentage of AUC values from the null distribution that were equal to or higher than the median AUC from the actual classification was taken as the p-value.

Results

See Table 3.

Full models

The *full models* performed successful classifications for all three input sets, namely *human languages*, *non-human vocalizations*, and when the two input sets were *combined*. These models achieved significance ($ps < .001$), with median AUCs of .999, and .999, and .982 respectively.

This is consistent with the possibility that there are shared surface properties available in the input sets, and that these are, in principle, available for infants to identify candidate links.

Table 3: Classification results, expressed as median area-under-the-curve (AUC) values for each input set and each model. ***$p < .001$, **$p < .01$, *$p < .05$

	AUC	Full	Spectral	Rhythmic	Inton'l
Input Sets:	*Human languages*	***.999	***1.000	***.916	*.641
	Non-human vocalizations	***.999	***.993	***.999	***.735
	Combined	***.982	***.972	***.953	.560

Spectral envelope models:

The *spectral envelope* models performed successful classifications when either *human languages* or *non-human vocalizations* were taken as input. These models achieved significance ($ps < .001$), with median AUCs of 1.000 (rounded up from .9996) and .993, respectively.

Successful classification results on *human languages* as input was not predicted; we expected that *spectral envelope* features, representing speech segments, would not play a role in identifying a candidate link to cognition in *human language*. Successful classification results on *non-human vocalizations* are consistent with our prediction that *spectral envelope* features, representing vocal tract configurations of animal species, may play a role in identifying candidate links to cognition in *non-human vocalizations*.

The *spectral envelope* models also successfully performed classifications on the *combined* input set ($p < .001$), with a median AUC of .972. This is consistent with the possibility that *spectral envelope* features, shared by human languages and non-human vocalizations, may identify them as candidate links to cognition.

Rhythmic models:

The *rhythmic* models performed successful classifications when either *human languages* or *non-human vocalizations* were taken as input. The two models achieved significance ($ps < .001$), with median AUCs of .916, and .999, respectively. This is consistent with the predictions that rhythm may play a role in identifying a candidate link to cognition in human languages and non-human vocalizations respectively.

The *rhythmic* models also successfully performed classifications on the *combined* input set ($p < .001$), with a median AUC of .953. This is consistent with the possibility the possibility that *rhythmic* features, shared between human languages and non-human vocalizations, may identify them as candidate links to cognition.

Intonational models:

The *intonational* models also performed statistically significant classifications when *human languages* or *non-human vocalizations* were taken as input ($ps < .05$), although by qualitative comparison, the AUCs of .641 on *human languages and*

.736 on *non-human vocalizations* are lower than its counterparts in *full*, *spectral envelope* and *rhythmic* models.

In contrast, the intonational model on combined input with a median AUC of .560 was not significant ($p > .05$). The lack of common surface properties available in intonational features in combined input is consistent with the possibility that the surface intonational properties, that contribute to the identification of candidate links from linguistic vocalizations, differ from those for non-human vocalizations.

Discussion

The current results were designed to bring the power of an supervised ML approach to the problem of specifying what acoustic information, might be present in the input of human and non-human vocalizations to successfully classify those vocalizations that either do or do not support object categorization in very young infants. Adopting the vantage point of 4-month-old English-acquiring infants, our models were designed to consider three well-documented classes of acoustic information, including *spectral envelope* features, *rhythmic* features, and *intonational* features. At issue was **a)** which models, if any, could be trained to make classifications that successfully distinguish vocalizations that support cognition from those that do not, and **b)** which class(es) of acoustic features support that classification.

Full models

Consider first, the performance of *full models*, which used all acoustic features (i.e. *spectral envelope*, *rhythmic*, and *intonational* features) to perform classifications. These models successfully classified vocalizations that support cognition and those that do not whether the model was trained on *human languages*, *non-human vocalizations*, and the two *combined* as input.

The success of the full models (using all three types of input) provides evidence that there are shared acoustic features present in the surface of human language and non-human linguistic vocalizations which, *in principle*, are available for infants to potentially establish certain signals as candidate links to cognition.

To further understand the nature of these shared acoustic features, we now summarize the respective models based on three well-defined classes of acoustic features (*spectral envelope*, *rhythmic*, or *intonational*). These three models highlight the specific shared acoustic properties that may potentially index candidate links for human languages, non-human vocalizations, as well as the two combined. We now turn to those more specific models.

Specific models

Spectral envelope models yielded robust classifications for *human languages* and *non-human vocalizations*, both individually and when *combined*.

The *spectral envelope* model's success with *human languages* as input was unexpected. Segmental information in human speech (e.g., consonants, vowels, syllables) is richly represented in the spectral envelope, but existing evidence suggests that infants younger than 6 months may not yet use this information in tuning to native segmental contrasts (Kuhl & Rivera-Gaxiola, 2008; Werker, 2018). At issue, then, is whether infants as young as 4 months use segmental cues to identify candidate links between language and cognition. Certainly, there is currently no evidence to suggest that the *spectral envelope* features identified by the ML algorithms are used, in practice, by 4-month-old infants. Instead, this ML evidence reveals only that *spectral envelope* features, whenever they do become available to infants, may be among the properties infants use to identify candidate links between human language and cognition.

As predicted, the *spectral envelope* model on *non-human vocalizations* yielded successful classification, suggesting that *spectral envelope* features are instrumental in identifying which non-human vocalizations are candidate links to cognition. This outcome is consistent with the hypothesis that links to cognition in human infants may be restricted to those produced by other primates, whose vocalizations are most similar to our own, both articulatorily and perceptually. In term of physiology, humans and non-human primate vocalize through a larynx, while birds do so through a syrinx. Perhaps the acoustic consequences of this physiologic difference are represented in *spectral envelope* features.

The success of the *spectral envelope* model on *combined* inputs is surprising. It opens new avenues for investigation. For example, in future work, it will be important to assess whether lemur vocalizations have the same facilitative effect on categorization in infants acquiring languages, like Cantonese, with segmental inventories (hence *spectral envelope* features) that differ systematically from those of English.

Rhythmic feature models were also robust across classifications on *human languages* and *non-human vocalizations*, both individually and when *combined*. This is consistent with the hypothesis that rhythm, is not only essential to infants' earliest speech perception (Christophe et al., 2001; Dehaene-Lambertz, Dehaene, & Hertz-Pannier, 2002; Gleitman & Wanner, 1982; Kuhl & Rivera-Gaxiola, 2008; Werker, 2018), but also to identifying which signals are candidate links to cognition. This evidence, coupled with the putative fundamental role of rhythm in infants' early language processing at the neural level (Goswami, 2019), raise a provocative possibility: that rhythmic properties of the infant's native language form a template for identifying candidate links to cognition. For example, linguistic inputs that abide to the rhythmic template of their native language would be linked to cognition. German, given its similarity in speech rhythm to English, can thus be linked to cognition by native English-listening infants. Given the domain generality of rhythm (Ravignani et al., 2019), rhythm may also be instrumental for non-human vocalizations in their links to cognition, as suggested by results of the classifications on *non-human vocalizations*. In particular, it is possible that the postulated rhythmic template for the language-cognition link may also operate over non-linguistic vocalizations, which could be potentially one explanation for the robust classification on *combined* inputs based on *rhythmic* features.

Intonational features constitute another component of prosody besides rhythm (Nooteboom, 1997). *Intonational models*, like *rhythmic* models, successfully classified vocalizations that support cognition from those that do not, using either *human languages* or *non-human vocalizations* as input.

However, when both human languages and non-human vocalizations were *combined* as input, the *intonational* model failed. That is, the model failed to identify common surface intonational properties by which candidate links to cognition from the combination of linguistic and non-human vocalizations, suggesting that at least some surface intonational properties are different across linguistic and non-human vocalizations.

This outcome is consistent with the hypothesis that infants' initial forays in identifying which signals are candidate links to cognition may follow two routes, one governing the links from language and another governing the candidate links from non-human vocalizations (Ackermann, Hage, & Ziegler, 2014; Owren, Amoss, & Rendall, 2011; Perszyk & Waxman, 2016, 2018). The linguistic and non-linguistic routes may take as input different sets of acoustic features in identifying which signals are candidate links to cognition. Therefore, one interpretation of the current results is that different sets of features related to intonation may be taken by the two routes as input, potentially bolstered by factors such as the different pitch-related properties (e.g., pitch range) produced by human and non-human vocal apparatuses (Charlton & Reby, 2016).

Certainly, these routes are *not entirely* distinct. Common surface *spectral envelope* and *rhythmic* features were identified by the models with *combined* input. Therefore, in principle, it is possible that the same set of *spectral envelope* and *rhythmic* features, but not *intonational* features, are utilized in the linguistic and non-linguistic routes in identifying which signals are candidate links to cognition.

We acknowledge a caveat of our current *intonational* models: even for statistically significant classifications on *human languages* and *non-human vocalizations*, they yielded surprisingly low classification performance. This may reflect a limitation of our corpus. After all, the f0 contour, although widely used to represent speech intonation, was the only intonational feature we were able to capture in the current models. As a result, we may have failed capture the dynamic intonational properties that infants perceive. The limited amount of information represented in the f0 contour as compared to spectral envelope and rhythmic features may also have hindered classification performance from a computational perspective. Addressing this question will require additional work that incorporates a broader and more dynamic set of intonational measures.

Limitations

Anchored by existing behavioral evidence on 3- to 4- months-old English-acquiring infants (Ferry et al., 2013; Perszyk & Waxman, 2019; Woodruff Carr et al., 2021), our study was limited to only five types of behaviorally attested vocalizations in the modeling. Therefore, our robust classification may reflect intrinsic acoustic differences in vocalizations across the specific non-human species or human languages tested in the models, rather than broader systematic differences between vocalizations which do and do not support cognition. Addressing this limitation will require the expansion of the variety of languages and non-human vocalizations attested behaviorally on infants, which would allow future work to better delineate the boundary conditions of vocalizations that do and do not support cognition.

Conclusion

The success of our ML models on individual classes of acoustic features suggests that there are *spectral envelope* and *rhythmic* features in the input of human language and of non-human linguistic vocalizations that identify certain signals as candidate links to cognition. This *in principle* evidence, important in itself, suggests that there may be evidence on the surface of vocalizations that lead infants to identify certain signals as being candidate links to cognition. This new evidence also suggest that infants' precocious establishment of a language-cognition link may be subserved by their perceptual sensitivity to the spectral envelope and rhythmic properties of sounds.

Acknowledgments

This work was supported by National Institute of Health (NIH) Grant R01HD083310 to S.R.W. We thank Patrick Wong, Luchang Wang, Sam Mehr, Chris Mercer, and Baby Speech Lab at the University of Konstanz for providing us with vocalization samples. We also thank Kali Woodruff Carr and Elena Luchkina for providing comments on the study and edits to the paper.

References

Ackermann, H., Hage, S. R., & Ziegler, W. (2014). Brain mechanisms of acoustic communication in humans and nonhuman primates: An evolutionary perspective. *Behavioral and Brain Sciences*(6), 529–546.

Andén, J., & Mallat, S. (2014). Deep scattering spectrum. *IEEE Transactions on Signal Processing*, *62*(16), 4114–4128.

Charlton, B. D., & Reby, D. (2016). The evolution of acoustic size exaggeration in terrestrial mammals. *Nature Communications*, *7*(1), 1–8.

Chong, A. J., Vicenik, C., & Sundara, M. (2018). Intonation plays a role in language discrimination by infants. *Infancy*, *23*(6), 795–819.

Christophe, A., Mehler, J., & Sebastián-Gallés, N. (2001). Perception of prosodic boundary correlates by newborn infants. *Infancy*, *2*(3), 385–394.

Dehaene-Lambertz, G., Dehaene, S., & Hertz-Pannier, L. (2002). Functional neuroimaging of speech perception in infants. *science*, *298*(5600), 2013–2015.

Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews*, *81*, 181–187.

Ferry, A. L., Hespos, S. J., & Waxman, S. R. (2010). Categorization in 3-and 4-month-old infants: an advantage of words over tones. *Child development*, *81*(2), 472–479.

Ferry, A. L., Hespos, S. J., & Waxman, S. R. (2013). Nonhuman primate vocalizations support categorization in very young human infants. *Proceedings of the National Academy of Sciences*, *110*(38), 15231–15235.

Filippi, P. (2016). Emotional and interactional prosody across animal communication systems: a comparative approach to the emergence of language. *Frontiers in Psychology*, *7*, 1393.

Gelman, S. A. (2004). Psychological essentialism in children. *Trends in cognitive sciences*, *8*(9), 404–409.

Gleitman, L., & Wanner, E. (1982). The state of the state of the art. In E. Wanner & L.Gleitman (Eds.), *Language acquisition: The state of the art* (p. 3-48). CambridgeUniversity Press.

Goswami, U. (2019). Speech rhythm and language acquisition: an amplitude modulation phase hierarchy perspective. *Annals of the New York Academy of Sciences*, *1453*(1), 67–78.

Kotz, S., Ravignani, A., & Fitch, W. (2018). The evolution of rhythm processing. *Trends in cognitive sciences*, *22*(10), 896–910.

Kuhl, P., & Rivera-Gaxiola, M. (2008). Neural substrates of language acquisition. *Annu. Rev. Neurosci.*, *31*, 511–534.

Laboratory of Vocal Learning at Hunter College. (2015). *Zebra finch song library 2015.* (data retrieved from `http://ofer.sci.ccny.cuny.edu`)

Mercer, C. (2012). *The audible phylogeny of lemurs.* Digital audio collection in posession of the author.

Miller, G. A. (1990). The place of language in a scientific psychology. *Psychological Science*, *1*(1), 7–14.

Moser, C. J., Lee-Rubin, H., Bainbridge, C. M., Atwood, S., Simson, J., Knox, D., … others (2020). Acoustic regularities in infant-directed vocalizations across cultures. *bioRxiv*.

Murphy, G. (2004). *The big book of concepts.* MIT press.

Nooteboom, S. (1997). The prosody of speech: melody and rhythm. *The handbook of phonetic sciences*, *5*, 640–673.

Owren, M. J., Amoss, R. T., & Rendall, D. (2011). Two organizing principles of vocal production: Implications for nonhuman and human primates. *American Journal of Primatology*, *73*(6), 530–544.

Peña, M., Pittaluga, E., & Mehler, J. (2010). Language acquisition in premature and full-term infants. *Proceedings of the National Academy of Sciences*, *107*(8), 3823–3828.

Perszyk, D. R., & Waxman, S. R. (2016). Listening to the calls of the wild: The role of experience in linking language and cognition in young infants. *Cognition*, *153*, 175–181.

Perszyk, D. R., & Waxman, S. R. (2018). Linking language and cognition in infancy. *Annual review of psychology*, *69*.

Perszyk, D. R., & Waxman, S. R. (2019). Infants' advances in speech perception shape their earliest links between language and cognition. *Scientific reports*, *9*(1), 1–6.

Ravignani, A., Dalla Bella, S., Falk, S., Kello, C., Noriega, F., & Kotz, S. (2019). *Evolution of speech rhythm: a cross-species perspective* (Tech. Rep.). PeerJ Preprints.

Tilsen, S., & Arvaniti, A. (2013). Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages. *The Journal of the Acoustical Society of America*, *134*(1), 628–639.

Wang, L., Kalashnikova, M., Kager, R., Regine, L., & Patrick, C. (2021). Lexical and prosodic pitch modifications in cantonese infant-directed speech. *Journal of Child Language*, 1–27.

Werker, J. F. (2018). Perceptual beginnings to language acquisition. *Applied Psycholinguistics*, *39*(4), 703–728.

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.

Woodruff Carr, K., Perszyk, D. R., & Waxman, S. R. (2021). Birdsong fails to support object categorization in human infants. *Plos one*, *16*(3), e0247430.

Zahner, K., Schönhuber, M., Grijzenhout, J., & Braun, B. (2016). Konstanz prosodically annotated infant-directed speech corpus (KIDS corpus). In *Speech prosody 2016* (pp. 562–566).