Comparison of speech and music input in North American infants' home environment over the first two years of life

Lindsay Hippe[1,2], Victoria Hennessy[1], Naja Ferjan Ramirez[1,3], T. Christina Zhao[1,2]

1. Institute for Learning and Brain Sciences, University of Washington, Seattle, WA, U.S.A.
2. Department of Speech and Hearing Sciences, University of Washington, Seattle, WA, U.S.A.
3. Department of Linguistics, University of Washington, Seattle, WA, U.S.A.

Data Availability Statement: data in .csv formats along with analysis code are available on OSF at Hippe, L., & Zhao, T. C. (2023, December 15). Tots and Tunes. Retrieved from osf.io/84rt9.

Conflict of Interest: The authors report no conflict of interest.

Institute Review Board Approval: The research reported in this publication received IRB approval under STUDY00016647 (Approval date: 11/23/2022)

Acknowledgements: Research reported in this publication was supported by the National Institute On Deafness And Other Communication Disorders of the National Institutes of Health under Award Number R01DC020419 (PI: Zhao). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This publication uses data generated via the *Zooniverse.org* platform, development of which is funded by generous support, including a Global Impact Award from Google, and by a grant from the Alfred P. Sloan Foundation. We thank Aditi Marehalli for annotating the dataset as the gold standard coder and Peter Mason for his substantial support with Zooniverse.

Author Contribution:
Conceptualization: CZ; Data curation: NFR; Formal analysis: LH; Methodology: LH, TH, CZ; Software: TH; Visualization: LH, TH; Writing–original draft**:** LH, CZ; Writing–review & editing: LH, TH, CZ, NFR

Corresponding Author: T. Christina Zhao, zhaotc@uw.edu, University of Washington BOX 357988, Seattle, WA, 98195

Abstract

Infants are immersed in a world of sounds from the moment their auditory system becomes functional and their experience with the auditory world shapes how their brains process sounds in their environment. Speech and music are two dominant auditory signals infants hear across cultures of the world. Decades of research have repeatedly shown that both quantity and quality of speech input play critical roles in infant language development. Less is known about the music input infants receive in their environment. This study is the first to compare music input to speech input across infancy, by analyzing a longitudinal dataset with daylong audio recordings collected in English-speaking naturalistic homes environments when the infants were 6, 10, 14, 18 and 24 months old. Using a citizen science approach, 643 naïve listeners annotated 12000 short snippets (10s) randomly sampled from the recordings using Zooniverse, an online citizen science platform. Results show that infants overall receive significantly more speech input than music input and the gap widens as the infants get older. Across all ages, infants experienced more music through electronic device than an in-person source; this pattern was reversed for speech. The percentage of music input intended for infants remained the same over time while that percentage for speech significantly increased. We propose possible explanations for the minor role of music compared to speech input observed in the present (North American) dataset and discuss future directions. We also discuss the opportunities in using a citizen science approach in analyzing large audio datasets.

Keywords: Auditory Environment, Speech Input, Music Input, LENA, Infancy, Citizen Science

Introduction

Infants are immersed in a world of sound from the moment their auditory system becomes active around the beginning of the third trimester (Hepper & Shahidullah, 1994). By the time they are born, they have already learned many things about their auditory environment, including their mothers' voice (DeCasper & Fifer, 1980), their mothers' language (Moon et al., 1993), as well as the lullabies they heard while in the womb (Partanen et al., 2013). This learning continues throughout early development and is constantly shaped by infants' auditory experience. Therefore, to better understand auditory learning in early development, it is crucial to better understand the auditory environment of infants.

Speech and music are two large components of human soundscape, universal across cultures (Hilton et al., 2022; Mehr et al., 2019). When interacting with infants, adults universally alter their communication style to a special style commonly described as infant-directed (ID) speech and song/singing (Hilton et al., 2022). Acoustic characteristics of ID vocalizations have been described extensively in the literature, particularly for speech. For example, ID speech is generally characterized by its higher pitch, expanded pitch range, exaggerated and distinctive vowels, and slower speaking rate (Cox et al., 2023; Hilton et al., 2022; Kuhl et al., 1997). On the other hand, ID songs feature reduced intensity and acoustic roughness with more energy at the lower frequency as well as more exaggerated rhythm (Hilton et al., 2022; Nakata & Trehub, 2011; Trainor et al., 1997). However, it is critical to note that there are many other dimensions to ID communication than acoustic modifications, such as using simplified vocabulary and syntactical structure (Genovese et al., 2020) and higher level of emotional expression (Hennessy & Zhao, 2023; Nguyen et al., 2023). Cross-culturally, infants display a preference for both ID speech and ID song over adult-directed speech and song (Byers-Heinlein et al., 2021;

ManyBabies Consortium, 2020; Trainor, 1996). These findings underscore the universal appeal of both ID speech and ID song, emphasizing the importance of studying them within the context of infants' naturalistic acoustic environments. Interestingly, when compared against each other, infants exhibit a preference for infant-directed song over speech (Nakata & Trehub, 2004; Tsang et al., 2017).

Over the last few decades, speech input in infants' auditory environment has been studied to a much greater extent than song/singing. Many studies have demonstrated repeatedly that both the quantity and quality of speech input in infancy, particularly the specialized form of ID speech, has profound long-term impact on infants, particularly on their language development (Cartmill et al., 2013; Rowe, 2012). Earlier studies used methodologies such as annotating short videos of mother-infant free play in lab (Dave et al., 2018) or in home (Rowe, 2012). These methods were limited in the amount of data and contexts they could capture. In more recent years, researchers have further homed in on examining naturalistic speech input in infants' home environment by collecting daylong audio recordings. Technologies such as the Language ENvironment Analysis (LENA) system utilize a small wearable recording device to capture infants' auditory environment throughout the day and generate recordings up to 16 hours in length. Indeed, quality ID speech annotated within such daylong recordings has been shown to be correlated with concurrent and later language skills (Ramirez-Esparza et al., 2014; Weisleder & Fernald, 2013) as well as children's brain structure and function (Huber et al., 2023; Romeo, Leonard, et al., 2018; Romeo, Segaran, et al., 2018). A recent intervention study solidified the causal effect of ID speech by demonstrating that parents who received early language coaching produced a higher quantity and quality of ID speech input compared to a control group, and that infants who received such high-quality speech input showed enhanced language skills (Ferjan Ramírez et al., 2020).

By contrast, only a handful of studies so far have examined musical input in infants' auditory environment, with the main approach being qualitative and quantitative parental report (Ilari, 2005; Politimou et al., 2018; Yan et al., 2021; Young, 2008). Notably, Politimou and colleagues recently developed and validated a comprehensive survey named Music@Home that quantitatively captures distinct aspects of home music environments (e.g., parental belief, parent singing etc.) (Politimou et al., 2018). Interestingly, music input assessed by Music@Home has been shown to correlate with concurrent gesture and word comprehension in younger infants (Papadimitriou et al., 2021). However, parents may not be the most reliable source of information as they have been shown to overestimate the amount of talking and singing to their children (Costa-Giomi & Benetti, 2017; Richards et al., 2017). So far, only a few studies have used naturalistic daylong recordings to assess infants' musical environment (Costa-Giomi & Benetti, 2017; Mendoza & Fausey, 2021). Mendoza and Fausey (2021) used LENA recordings supplemented by manual annotation to quantitatively assess the amount of musical input in English speaking North American families with infants (n = 35) aged between 6-12 months. Singing was found to be present in more than half of everyday music for infants, and instrumental music was present in more than three-quarters. Only one third of the music was from live sources, while three quarters was from recorded sources. Additionally, infants tended to hear certain songs (e.g., favorite nursery rhymes) and/or voices (e.g., mother) more frequently than others. Critically, and unlike the present study, Mendoza and Fausey relied on labor-intensive manual annotation of the entire dataset through a large group of trained undergraduate students (N= 38) and focused solely on infants' musical input.

To date, no known study has directly compared speech versus music input in infants' naturalistic auditory environment. Furthermore, it is unknown how this speech versus music input may change over the course of infant development. The current study addresses this important gap by analyzing an existing longitudinal LENA dataset collected longitudinally in North American English-speaking families when the infants were 6, 10, 14, 18 and 24 months. We describe speech and music input side by side throughout infancy and address three main questions: First, we examine the total amount of speech input in comparison to music input over the first two years of life. Second, given the increasing prevalence of electronic devices in recent years, we further examine the source of input in speech versus music (Mendoza & Fausey, 2021; Young, 2008). Third, we delve deeper into each input type and examine whether it was intended for infants or not (i.e., was the speech or music infant-directed).

To extract quantitative measure of speech and music input (e.g., amount of speech in infants' environment), we developed a citizen-science method for LENA data annotation. This approach was chosen for several reasons. First, LENA's automated software can conduct rudimentary analysis of audio data with a focus on speech input (e.g., adult word count, child word count, silence). Automatic detection of more fine-grained characteristics (e.g., infant-directedness, source of input) and automatic characterization of sounds outside of speech domain (e.g., music) is largely unavailable or unreliable. Second, it was feasible for the current study to adopt the citizen-science approach to produce high quality data given it met the following conditions (Kosmala et al., 2016). 1. Our annotation categories are simple with limited choices, thus do not require additional training for participants. Indeed, it has been shown previously that naïve listeners can distinguish between audio recordings of infant- and adult-directed speech and songs reliably at a level more accurate than chance (Hilton et al., 2022). 2. A majority voting procedure was implemented to derive the final annotation, enhancing its accuracy. 3. Reliability was examined and validated between annotations derived from the citizen science approach vs. a trained coder. Specifically, short snippets of audio recordings were randomly sampled from the daylong recordings, and each was annotated by multiple naïve listeners. The majority voting method was employed to derive the final annotation for these snippets. The count of snippets with final annotations under each category (e.g., total speech, total music, infant-directed speech, infant-directed music etc.) were taken as dependent measures (details in Methods section).

Methods

Stimulus

The stimuli for the current annotation experiment were extracted from an existing dataset consisting of daylong auditory recordings of infants' sound environment made with Language ENvironment Analysis (LENA) recording devices (Ferjan Ramírez et al., 2019, 2020; Huber et al., 2023). In the original study, infants wore the small LENA recording device for two days to record their naturalistic sound environment for up to 16 hours per day at each recording age. They were recorded 5 times longitudinally at 6, 10, 14, 18 and 24 months of age, on average within three days of the target date. All families were monolingual English-speaking. Some of the families received parent coaching on infant-directed speech production (Intervention Group) while the others did not (Control Group). All infants were born full term (within ±14 days of due date), of normal birth weight (6–10 lbs.) and had no major birth or postnatal complications. In the current study, we utilized only the Control Group dataset and all infants had complete sets of LENA recordings (i.e., N =24 with 12 male infants, two-day recordings at each of the 5 ages).

The LENA Advanced Data EXtractor (ADEX) tool was first used to divide each daylong recording into five-minute segments, with each segment containing basic information on a variety of automatically derived variables (e.g., Female Adult, Male Adult, Key Child, Other Child, Silence, etc.). We examined the distribution of the Silence variable across all infants and observed a bimodal distribution (see Figure S1) where the top 10 percent of the 5-minute segments ranked by the amount of Silence are predominantly silent (mean = 239.00, std = 88.15). Therefore, we dropped the top 10 percent of 5-minute segments with the highest amount in Silence variable from each recording. From the remaining of each recording, we randomly selected 50 10-second snippets of audio, resulting in 100 10-second samples per subject per age. The length of the snippet (10 seconds) was chosen to mitigate the inclusion of potential confidential information, while still having enough information for participants to annotate. In total, 12000 snippets across 5 ages (i.e., 2400 snippets per age) were used as the stimuli for the current annotation experiment.

Participants

We recruited participants (N = 643) for the current annotation experiment through the Online Research Pool Program (ORPP) in the Department of Psychology at the University of Washington. Students who participated in the experiment received extra credits in their Psychology courses. No other information was obtained from these participants. The annotation experiment procedure was approved by the Institute Review Board at the University of Washington.

Experimental Procedure

The current annotation experiment was conducted on the Zooniverse, an online platform specialized for conducting citizen science research. Critically, Zooniverse is capable of hosting large datasets (up to 1 MB per individual file) and allows for flexible project design through a client-oriented Python module that provides high-level access to the Zooniverse API

(Application Programming Interface). For the current experiment, audio snippets were randomly selected from the stimulus set and presented to participants on the project's Zooniverse page. Participants were asked to listen to either 50 or 100 snippets and answer a series of questions about each 10-second snippet (i.e., providing annotation, see Figure 1 and paragraph below for details). We utilized Caesar, an advanced Zooniverse tool that allows tracking and aggregation of participant responses in real-time, to customize the circulation of audio snippets based on live participant selections. Individual snippets ceased to be presented to participants (i.e., were retired) once three different participants selected the None/Other option or at least five different participants voted.

Two levels of randomization were used for stimulus presentation. First, sequence of age was randomized (i.e., 6, 14, 18, 24, 10). Within each age, Zooniverse randomly selected from stimuli that have not yet been retired each time. Specifically for annotation, there were three main questions asked (Figure 1). In the first question, all participants answered the question "What do you hear in this clip?" The choices given were (1) speech produced by someone older than two years (Speech from this point), (2) music or singing (Music from this point), (3) speech produced by someone older than two years AND music/singing (Speech AND Music from this point), and (4) none of the above/other (None from this point). Annotation ended for the snippet if participants chose the 4th option (i.e., None). Participants continued to answer the other two main questions if they chose options 1-3. Further, if they chose the 3rd option (Speech AND Music), they answered the following questions twice, once for speech and once for music. In the second main question, participants answered "What type of (Speech/Music) do you hear?" The choices given were (1) in-person (Speech/Music), (2) (Speech/Music) through an electronic device, and (3) both of these. In the third main question, participants answered “Is the (Speech/Music) directed to a baby?” The choices given were (1) yes and (2) no.

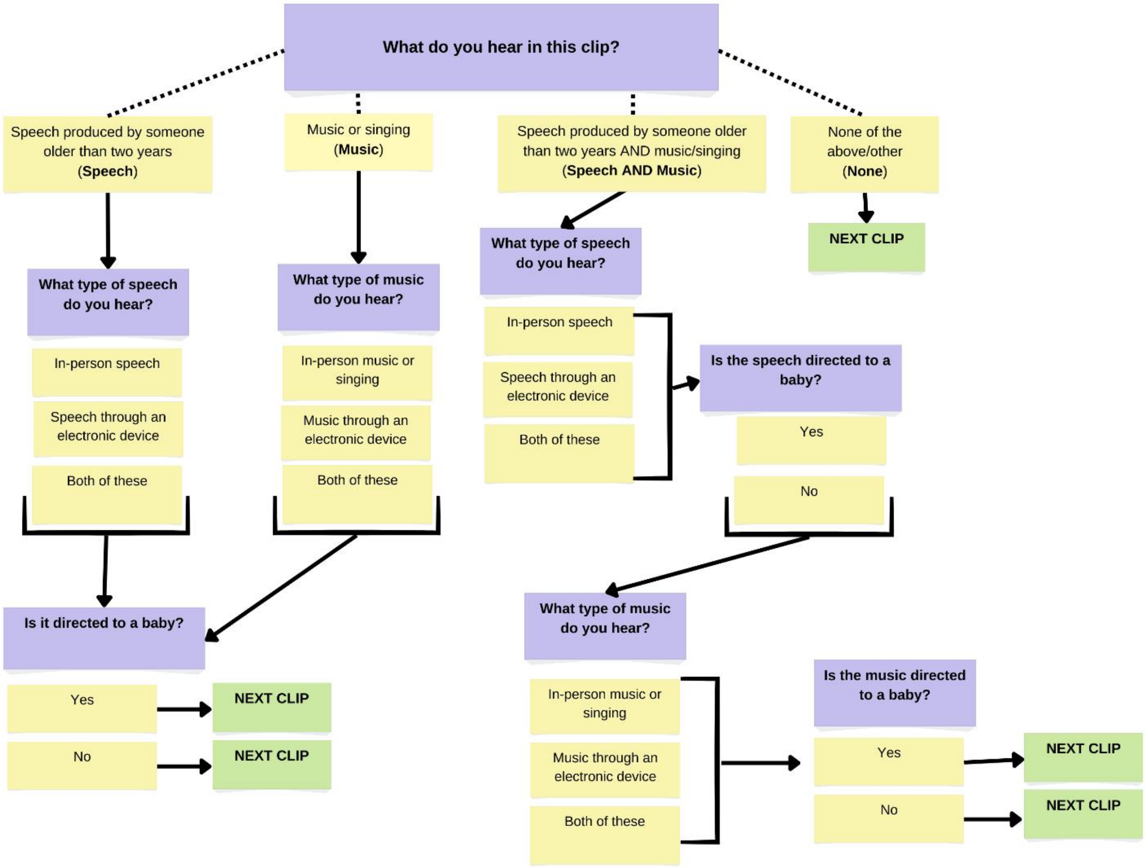
What do you hear in this clip?
Speech produced by someone older than two years (Speech)
Music or singing (Music)
Speech produced by someone older than two years AND music/singing (Speech AND Music)
None of the above/other (None)
NEXT CLIP
What type of speech do you hear?
In-person speech
Speech through an electronic device
Both of these
What type of music do you hear?
In-person music or singing
Music through an electronic device
Both of these
What type of speech do you hear?
In-person speech
Speech through an electronic device
Both of these
Is the speech directed to a baby?
Yes
No
Is it directed to a baby?
Yes
No
NEXT CLIP
NEXT CLIP
What type of music do you hear?
In-person music or singing
Music through an electronic device
Both of these
Is the music directed to a baby?
Yes
No
NEXT CLIP
NEXT CLIP

Figure 1. Illustration of the annotation question flow on Zooniverse.

Data Processing

Raw annotation data (.csv files) were exported from Zooniverse and were processed using in-house python scripts (Hippe & Zhao, 2023, Tots & Tunes: https://osf.io/84rt9/). Specifically, we first transformed the data structure such that each row contains the aggregated annotation data for each snippet from all participants who made annotation for that snippet. The percentage of votes for each option are calculated and stored for each question. For example, for the first question, the vector of [40, 0, 20, 40] would translate to 40% of votes for option 1

(Speech), 0% votes for option 2 (Music), 20% vote for option 3 (Speech AND Music), and 40% votes for option 4 (None). The summation of all numbers in this vector equals to 100.

Based on the raw vote percent, final annotations were derived for each snippet for the first question based on the following majority voting steps. The first pass ascertained whether one option had over 50 percent of the vote and, if so, marked that option as the final annotation for that snippet. For example, in the vector [20, 20, 0, 60], the "None" option has 60% of the vote, so the clip would be marked as "None." If the criterion for the first pass was not met, the second pass determined which option had the maximum percentage and marked that option as the final annotation for that snippet. For example, in the vector [40, 20, 20, 20], no option has over 50 percent of the vote, but the "Speech" option has the maximum percentage, so the corresponding clip would be marked as "Speech." The data went through the third pass if there was a tie for the maximum percentage. For example, in the vector [40, 0, 20, 40], two options received 40 percent of the vote, the maximum percentage. To address this, the value for the 3rd option (Speech AND Music) was added to the values of option 1 (Speech) and option 2 (Music), respectively. In this case, the resulting vector becomes [60, 20, 20, 40] with option 1 marked as the final annotation. If no final annotation could be derived after the three criteria, the snippet was marked as "unresolvable" and was not included in further analyses. For example, a snippet with a vector of [20, 40, 0, 40] does not meet any of the criteria because there is a tie for the maximum percentage that cannot be resolved by the third pass, as no votes were received for option 3 (Both). The majority of decisions on snippets were resolved through the first pass with a small percentage unresolved (see Table S1).

To further examine the validity of this majority voting approach, we trained a research assistant to annotate all 2400 snippets in the 6-month dataset as the gold standard coder (20% of all data) and we examined the reliability between annotation from the trained coder vs. final annotation derived from majority voting approach. For the first question, Cohen's Kappa was calculated based on the confusion matrix (See Table S2) and demonstrated good reliability between the two annotation methods (Kappa = 0.73).

Based on the annotation for the first question, all snippets marked as "None" were excluded (average number of 'None' snippets for each age: 6 months: 924, 10 months: 958, 14 months: 882, 18 months: 830, and 24 months: 782). Only snippets with final annotation of Speech, Music, or Speech AND Music were entered into further majority voting to derive final annotation for the second and third questions. For the question regarding the source of the Speech/Music, the option with the maximum count was taken as the final annotation for that snippet. In the example of a vector of [60, 0, 20], the snippet will be marked as 'In-Person'. If a tie existed between the 'In-Person' and 'Device' options, then 'Both' was taken as the final annotation, such as in the example of a vector of [40, 40, 20]. If the tie was between the 3rd option and either of the first two options, the same strategy used for the first question was employed by adding the number for the 3rd option ('both') to each of the first two options. In an example of [20, 0, 20], the vector then becomes [40, 20, 20] with final annotation as 'In-Person'. For the question regarding the intended target (Infant Directed: yes vs. no), the option with the larger number of votes is taken as the final annotation. However, if a tie exists, the snippet becomes unresolvable for that question and is excluded from the final count (number of unresolved snippets for this question: 6 months: 99, 10 months: 89, 14 months: 99, 18 months: 127, and 24 months: 74).

Once every snippet received its final annotations and was aggregated by each infant at a specific age, we then further eliminated the 'Both' categories for the first question (i.e., Speech

AND Music/Sing) and the second question (i.e., In Person AND electronic device), by adding the number of snippets in that category (e.g., Speech AND Music/Sing) to the other two categories (e.g., Speech Category and Music/Sing Category). For example, if one infant at 6 months received final annotation for 50 snippets for Speech, 10 for Music, 3 for Speech AND Music, then the infants' final counts are 53 for Speech and 13 for Music. The count for the 2nd and 3rd questions were propagated correspondingly under Speech and Music. For example, of the 53 snippets in the total speech category, 37 were 'In person', 14 'Through electronic device' and 2 as in 'Both', then the final count for 'In-person Speech' is 39 and 'Speech through electronic device' is 16. Further, of the 53 snippets in the total speech category, 26 were annotated as 'directed to baby', 25 were 'not directed to baby' with 2 'unresolved'. In the end, 10 dependent measures were derived for each infant at each age: number of snippets classified as (1) total speech, (2) total music/sing, (3) in-person speech, (4) speech through an electronic device, (5) speech directed to a baby, (6) speech not directed to a baby, (7) in-person music/sing, (8) music/sing through an electronic device, (9) music/sing directed to a baby, and (10) music/sing not directed to baby.

Statistical Analysis

All statistical analyses were conducted using R and R Studio software (R Studio Team, 2020; R Core Team, 2020). To address our primary research question, a main linear mixed-effect model was used (lme4 package) where Age (6, 10, 14, 18, and 24 month), Type of Input (total speech vs. total music) and their interaction were entered as fixed factors. In addition, individual infants (intercept plus slope) were entered as a random factor. Post hoc pair-wise comparisons were conducted for fixed factors (lmerTest package). To address our secondary questions, two separate linear mixed-effect models were used for investigating the question regarding input source vs. recipient. In one model, Age, Type of Input (Speech vs. Music), Source of Sound (In Person vs. Through Electronic Device) and the interactions were entered as fixed factors and individual infants (intercept and slope) were entered as a random factor. Similarly, in the other model, Age, Type of Input (Speech vs. Music), Target of Sound (Infant-Directed vs. Non-Infant-Directed) and the interactions were modeled as fixed factors and individual infant (intercept and slope) was entered as a random factor. Post hoc pairwise comparisons were conducted for fixed factors. Lastly, exploratory correlation analyses were conducted between total speech and total music across participants at each age.

Results

Question 1: Infants receive significantly more speech input than music input and the gap widens with infant age.

The amount of total speech vs. total music input infants receive across ages can be visualized in Figure 2. The linear mixed-effect model output reveals a significant effect of Type of input (Speech vs. Music) (F = 2106.41, p <0.001) and a significant interaction between Age and Type of input (F = 5.88, p = 0.002). The effect of Age was not significant (F = 2.16, p = 0.08). Pairwise post hoc comparisons revealed that total music input remained stable across ages while total speech input increased with age (age 6: age 18, p = 0.009, age 10: age 24, p = 0.0002, age 14: age 24, p = 0.002, age 18: age 24, p = 0.03).

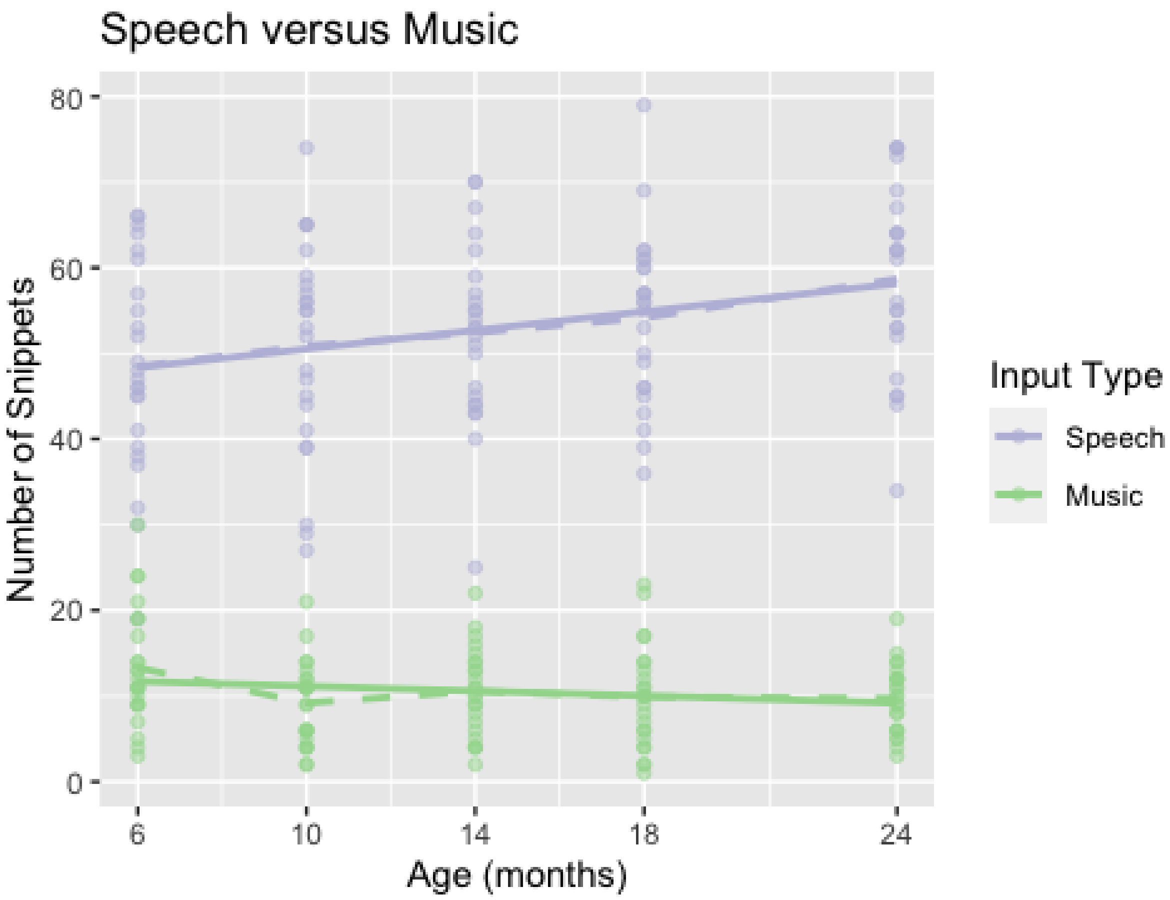

Figure 2 . Total number of snippets marked as speech versus total number of snippets marked as music, with a dashed line connecting the mean at each timepoint and a solid line representing a linear regression of the data.

Question 2: There is significantly more speech input from an in-person source than from an electronic source, while the pattern is reversed for music. Only speech input from an in-person source increased with infant age, while all other categories (speech from electronic source, music from in-person and electronic source) remained stable across ages.

The amount of input from an in-person vs. an electronic device for both speech and music that infants receive across ages is visualized in Figure 3. The linear mixed-effect model output reveals a significant effect of Type of input (Speech vs. Music) (F = 1435.69.41, *p* <0.001), a significant effect of Input Source (In Person vs. Electronic Device) (F = 955.10, *p* <0.001), a significant interaction between Age and Type of input (F = 4.07, *p* = 0.003), a significant interaction between Age and Input Source (F = 3.64, *p* = 0.006), a significant interaction between Input Type and Source (F = 1694.56, *p* <0.001) and critically, a significant 3-way interaction (F = 2.71, *p* = 0.03). Only the effect of Age was not significant (F = 1.27, *p* = 0.28). Pairwise post hoc comparisons revealed that music input remained stable across ages for both In-Person and Electronic Device sources (*p*s > 0.1). For speech, input from Electronic Device sources remained stable across ages (*p*s > 0.1) while input from In-Person Sources increased (6-10 mo, *p* = 0.025, 10-24 mo: *p* <0.001, 14-24mo: *p* <0.001, 18-24mo: *p* = 0.003).

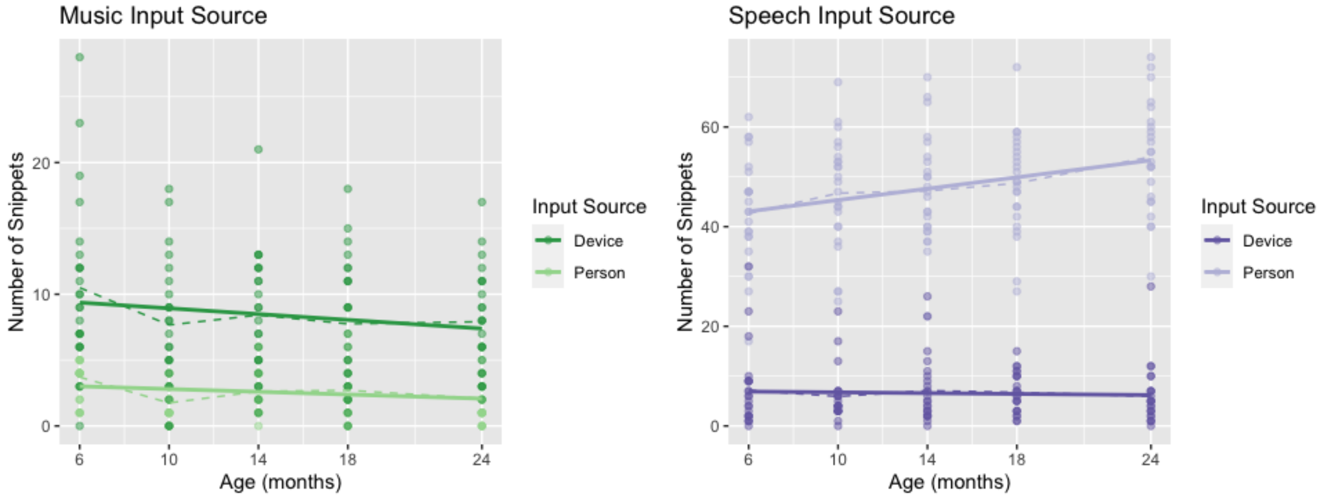

Figure 3 . Number of music (left) and speech (right) snippets marked as being produced by an electronic device (darker color) versus being produced by a person (lighter color).

Question 3: There is a smaller proportion of music input intended for infants and the proportion remains stable across ages. On the other hand, speech input intended for infants significantly increases with infant age, while speech input unintended for infants decreased.

The amount of input intended for the infant vs. not intended for the infant for both speech and music across ages is visualized in Figure 4. The linear mixed-effect model output reveals a significant effect of Type of input (Speech vs. Music) (F = 1164.56, *p* <0.001), a significant effect of Input Recipient (Infant vs. Not Infant) (F = 46.27, *p* <0.001), a significant interaction between Age and Type of input (F = 3.65, *p* = 0.006), a significant interaction between Age and

Input Recipient (F = 10.40, *p* < 0.001), and critically, a significant 3-way interaction (F = 9.96, *p* < 0.001). The effect of Age (F = 1.27, *p* = 0.28) and the interaction between Input Type and Input Recipient (F = 0.15, p = 0.70) were not significant. Pairwise post hoc comparisons revealed that music input remained stable across ages regardless of whether they are intended for the infant (*p*s > 0.1). For speech, input unintended for the infant was significantly lower at 24 months of age when compared to 6, 10 and 14 months (*p* = 0.003, 0.002, 0.003), while input intended for infants increased (6-14 mo, *p* = 0.008, 10-18 mo: *p* = 0.03, 14-24 mo: *p* <0.001, 18-24 mo: *p* < 0.001).

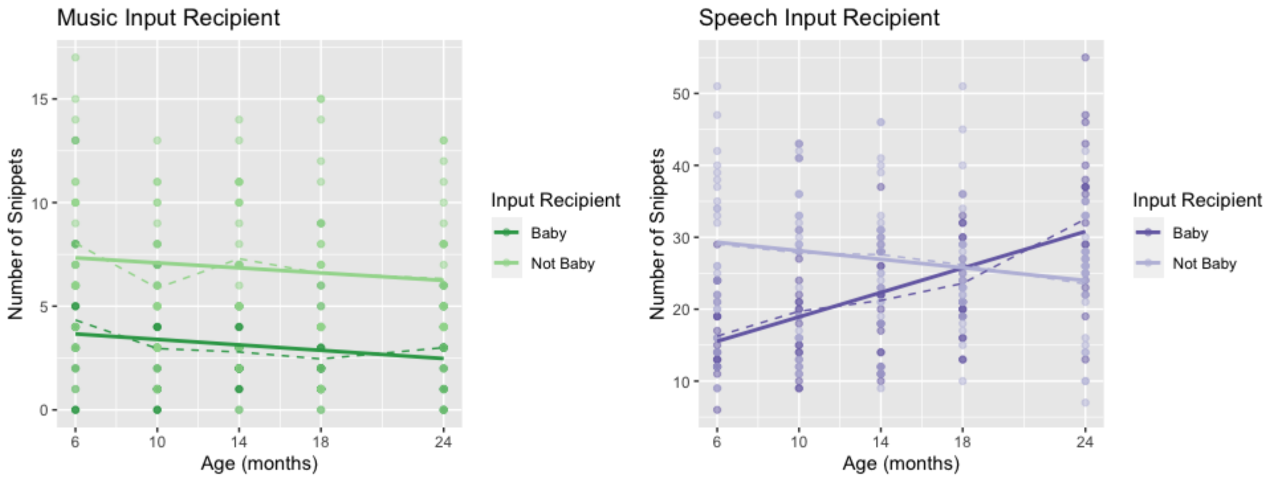

Figure 4 . Number of music (left) and speech (right) snippets marked as being directed to a baby (darker color) versus not directed to a baby (lighter color).

Exploratory question: no significant correlation between speech and music input across infants

The Pearson correlation coefficients were calculated between total speech and total music at each age are 0.47 (6 mo), 0.25 (10 mo), 0.08 (14 mo), 0.29 (18 mo), and 0.14 (24 mo). No *p* value was below 0.05 after adjusting for multiple tests.

Discussion

The present study quantitatively assessed speech and music input in North American infants' naturalistic auditory environment over the first two years of their lives. By utilizing a citizen science approach, short snippets randomly sampled from daylong LENA recordings were annotated by multiple naive listeners with final annotation derived through a majority voting procedure. Our results addressed three important research questions: (1) What is the total amount of speech versus music input across the first two years of life (i.e., in infancy)? Overall, we observed a significantly larger amount of speech input than music input across all 5 ages (6, 10, 14, 18, and 24 months) with the gap widening over time. (2) What is the distribution of in-person speech versus speech delivered through an electronic source? How does this compare to the distribution of in-person music versus music delivered through an electronic source? Examination of the source of the sound (i.e., in-person versus through an electronic device) revealed significantly more in-person than electronic speech and significantly more electronic than in-person music. Further, in-person speech was observed to increase over time while the other three categories remained relatively stable. (3) What is the distribution of infant-directed speech versus non-infant-directed speech? How does this compare to the distribution of infant-directed music versus non-infant-directed music? Our analysis on the intended target of the sound revealed that for speech, there was a cross-over in development when infants started to receive more ID speech than non-ID speech after 18 months of age. However, for music, these proportions remained unchanged and ID music was in the minority.

All of our findings regarding speech input are well aligned with the existing literature (Bergelson et al., 2023). The increasing amount of speech infants hear overall highlights its increasing dominance as an auditory input for infants. Particularly, the increase was largely driven by speech that is intended for infants and delivered in person. Our results on ID speech replicated a previous study demonstrating the simultaneous increase of ID speech and decrease of non-ID speech as infants get older (Bergelson et al., 2019). In our data, the cross-over (i.e., ID speech surpassing non-ID speech) happens around 18 months of age when conversation turn-taking increase drastically, supporting the idea that as infants become more communicative, their caregivers naturally start engaging them in more verbal and social communication, which may further propel infants' language acquisition (Ferjan Ramírez et al., 2021). Given the importance of social, in-person and infant-directed speech input (Ferjan Ramírez et al., 2020; Kuhl et al., 2003), it is reassuring to see that parents are highly engaged as infants' language skills develop.

However, our findings on music input, especially when compared to speech, are largely surprising to us. We found that within North American, English-speaking families, infants engage in few interactions with their caregivers in the form of live musical activities that would be specifically intended for them. This is surprising, given that singing behavior, and particularly infant-directed singing, has been documented as a universal phenomenon across cultures (Mehr et al., 2019, Hilton et al, 2022). Our findings suggest that, while these behaviors exist within most families, they are relatively rare during the day (number of infants with 0 snippet coded as infant-directed music at each age: 6 mo. (3), 10 mo. (3), 14mo. (1), 18mo. (1), 24mo. (3)). It is possible that this result is specific to this sample given the participating families were homogenous geographically and culturally (families in the Pacific Northwest of North America with mid to high socioeconomic status). It is also possible that, while music is universal, it is much less prevalent in infants' environment as it does not consistently and efficiently convey specific lexical information as speech (though see (Margulis et al., 2022). Future studies are warranted to examine the distribution of speech and music input across a wider range of cultures with different beliefs around the value of music and speech on infant development. Finally, a lack of correlation between speech and music input suggests that those two dimensions are likely contributing to infant development independently. Future research should consider both dimensions when studying infants' auditory environment.

Over the past decade, a growing and converging body of evidence has documented the multifaceted benefit of music experience for early development, such as speech and language learning (Zhao & Kuhl, 2016, Zhao, Llanos, Chandrasekaran & Kuhl, 2022). Furthermore, music experience has been demonstrated to foster social emotional connections between infants and their caregivers/their peers (Cirelli et al., 2014; Rabinowitch & Meltzoff, 2017), with social and infant-directed activities generating larger benefits than passive and non-infant directed ones (Gerry et al., 2012; Golinkoff et al., 2015; Kuhl, 2007). Given the minor role of music input observed in the current study, our findings point to significant needs and lots of potential for improving music input in North American infants' auditory environments, especially in the form of live, interactive, and infant-directed musical experience. Further studies might first gain deeper understanding of the reasons behind a general paucity of music in North American English-speaking families with young infants. Such insights will be necessary to help design parent-focused intervention methods to increase high-quality infant-directed musical activities to maximize the benefits from early music experiences.

Methodologically, we took a new citizen-science approach to annotate a large corpus of daylong LENA audio recordings. Traditionally, this type of work has been done by a group of trained research assistants in the lab, which is a reliable, but a labor and cost-intensive practice. We implemented a majority voting procedure that ensured the accuracy while reducing the cost of annotation. Indeed, a reliability analysis validated the annotation derived from the current approach as it was shown to be is highly consistent with annotation made by a trained coder.

Thus, we think this type of approach has lot of potential for analyzing LENA recordings in future work. Given that this is a relatively new approach, we also acknowledge several caveats that we learned over the course of the study. First, in this current study, we took a small random sample across 90% of the recordings to survey the amount of speech and music. While similar approaches are commonly used in LENA analyses, the small sample may not be completely representative of the entire daylong recording. We are developing better processing pipelines that can automatically detect silence and noise for exclusion and thus help better sample the audio recordings. Furthermore, different sampling approaches can be examined in the future as well (e.g., more targeted sampling vs. random sampling). Second, even though we used only a few relatively simple response categories (i.e. Speech, Music, None); ambiguity still existed. Examples of ambiguity included background speech that was unintelligible, or speech delivered in a 'rap-like' manner. Our trained coder has compiled such instances, which will help us to further improve our annotation instruction. Third, there is no ground truth regarding the intended recipient that can be uncovered solely from the LENA recording. In other words, we will never know for sure whether the input was indeed intended for the infant without more information (e.g., visual, context, input from the caregiver). Strong agreement among coders does not necessarily mean correct judgement. This issue is not specific to this study but applies to any annotation scheme that uses LENA recordings. Speech annotation can be slightly easier as some level of contextual information exists, even in short snippets (e.g., 'Here is your rubber duckie'). Annotating recipient for music, particularly instrumental music, is much harder. It is unclear what criteria were used by coders. It is possible that common knowledge about the music content was used for such judgements, for example, labeling well-known children's melodies (e.g., 'Wheels on the bus') as infant-directed. However, this approach may not work if coders are unfamiliar with habits of a specific household (e.g., using Taylor Swift songs for play or routine) or their culture (e.g., children's songs in other cultures), resulting in inaccurate/biased labeling. Future research is warranted to have better understanding of music in infants' environment, such as by examining what caregivers consider as infant-directed music.

In conclusion, the current study provided a first look at infants' speech and music input in their naturalistic home environment and observed large differences between the two domains. We believe these results demonstrate the need to further study music environments of infants across cultures in relation to their speech environment, and to further examine ways in which music input can be improved through intervention.

References

Bergelson, E., Casillas, M., Soderstrom, M., Seidl, A., Warlaumont, A. S., & Amatuni, A. (2019). What Do North American Babies Hear? A large-scale cross-corpus analysis. *Developmental Science*, *22*(1), e12724. https://doi.org/10.1111/desc.12724

Bergelson, E., Soderstrom, M., Schwarz, I.-C., Rowland, C. F., Ramírez-Esparza, N., R. Hamrick, L., Marklund, E., Kalashnikova, M., Guez, A., Casillas, M., Benetti, L., Alphen, P. van, & Cristia, A. (2023). Everyday language input and production in 1,001 children from six continents. *Proceedings of the National Academy of Sciences*, *120*(52), e2300671120. https://doi.org/10.1073/pnas.2300671120

Byers-Heinlein, K., Tsui, A. S. M., Bergmann, C., Black, A. K., Brown, A., Carbajal, M. J., Durrant, S., Fennell, C. T., Fiévet, A.-C., Frank, M. C., Gampe, A., Gervain, J., Gonzalez-Gomez, N., Hamlin, J. K., Havron, N., Hernik, M., Kerr, S., Killam, H., Klassen, K., … Wermelinger, S. (2021). A Multilab Study of Bilingual Infants: Exploring the Preference for Infant-Directed Speech. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920974622. https://doi.org/10.1177/2515245920974622

Cartmill, E. A., Armstrong, B. F., Gleitman, L. R., Goldin-Meadow, S., Medina, T. N., & Trueswell, J. C. (2013). Quality of early parent input predicts child vocabulary 3 years later. *Proceedings of the National Academy of Sciences*, *110*(28), 11278–11283. https://doi.org/10.1073/pnas.1309518110

Cirelli, L. K., Einarson, K. M., & Trainor, L. J. (2014). Interpersonal synchrony increases prosocial behavior in infants. *Developmental Science*, *17*(6), 1003–1011. https://doi.org/10.1111/desc.12193

Costa-Giomi, E., & Benetti, L. (2017). Through a baby's ears: Musical interactions in a family community. *International Journal of Community Music*, *10*(3), 289–303. https://doi.org/10.1386/ijcm.10.3.289_1

Cox, C., Bergmann, C., Fowler, E., Keren-Portnoy, T., Roepstorff, A., Bryant, G., & Fusaroli, R. (2023). A systematic review and Bayesian meta-analysis of the acoustic features of infant-directed speech. *Nature Human Behaviour*, *7*(1), 114–133. https://doi.org/10.1038/s41562-022-01452-1

Dave, S., Mastergeorge, A. M., & Olswang, L. B. (2018). Motherese, affect, and vocabulary development: Dyadic communicative interactions in infants and toddlers. *Journal of Child Language*, *45*(4), 917–938. https://doi.org/10.1017/S0305000917000551

DeCasper, A., & Fifer, W. (1980). Of Human Bonding: Newborns Prefer Their Mothers' Voices. *Science*, *208*(4448), 56.

Ferjan Ramírez, N., Hippe, D. S., & Kuhl, P. K. (2021). Comparing Automatic and Manual Measures of Parent-Infant Conversational Turns: A Word of Caution. *Child Dev*, *92*(2), 672–681. https://doi.org/10.1111/cdev.13495

Ferjan Ramírez, N., Lytle, S. R., Fish, M., & Kuhl, P. K. (2019). Parent coaching at 6 and 10 months improves language outcomes at 14 months: A randomized controlled trial. *Dev Sci*, *22*(3), e12762. https://doi.org/10.1111/desc.12762

Ferjan Ramírez, N., Lytle, S. R., & Kuhl, P. K. (2020). Parent coaching increases conversational turns and advances infant language development. *Proceedings of the National Academy of Sciences*, *117*(7), 3484–3491. https://doi.org/10.1073/pnas.1921653117

Genovese, G., Spinelli, M., Lauro, L. J. R., Aureli, T., Castelletti, G., & Fasolo, M. (2020). Infant-directed speech as a simplified but not simple register: A longitudinal study of lexical and syntactic features. *Journal of Child Language*, *47*(1), 22–44. https://doi.org/10.1017/S0305000919000643

Gerry, D., Unrau, A., & Trainor, L. J. (2012). Active music classes in infancy enhance musical, communicative and social development. *Developmental Science*, *15*(3), 398–407. https://doi.org/10.1111/j.1467-7687.2012.01142.x

Golinkoff, R. M., Can, D. D., Soderstrom, M., & Hirsh-Pasek, K. (2015). (Baby)Talk to Me: The Social Context of Infant-Directed Speech and Its Effects on Early Language Acquisition. *Current Directions in Psychological Science*, *24*(5), 339–344. https://doi.org/10.1177/0963721415595345

Hennessy, V., & Zhao, T. C. (2023). *Building the bond: The social-emotional role of infant-directed speech & song*. https://doi.org/10.31234/osf.io/sm4ux

Hepper, P. G., & Shahidullah, B. S. (1994). The development of fetal hearing. *Fetal and Maternal Medicine Review*, *6*(3), 167–179.

Hilton, C. B., Moser, C. J., Bertolo, M., Lee-Rubin, H., Amir, D., Bainbridge, C. M., Simson, J., Knox, D., Glowacki, L., Alemu, E., Galbarczyk, A., Jasienska, G., Ross, C. T., Neff, M. B., Martin, A., Cirelli, L. K., Trehub, S. E., Song, J., Kim, M., … Mehr, S. A. (2022). Acoustic regularities in infant-directed speech and song across cultures. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-022-01410-x

Huber, E., Corrigan, N. M., Yarnykh, V. L., Ferjan Ramírez, N., & Kuhl, P. K. (2023). Language Experience during Infancy Predicts White Matter Myelination at Age 2 Years. *The Journal of Neuroscience*, *43*(9), 1590–1599. https://doi.org/10.1523/jneurosci.1043-22.2023

Ilari, B. (2005). On musical parenting of young children: Musical beliefs and behaviors of mothers and infants. *Early Child Development and Care*, *175*(7–8), 647–660. https://doi.org/10.1080/0300443042000302573

Kosmala, M., Wiggins, A., Swanson, A., & Simmons, B. (2016). Assessing data quality in citizen science. *Frontiers in Ecology and the Environment*, *14*(10), 551–560. https://doi.org/10.1002/fee.1436

Kuhl, P. K. (2007). Is speech learning "gated" by the social brain? *Developmental Science*, *10*(1), 110–120. https://doi.org/10.1111/j.1467-7687.2007.00572.x

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, U., & Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science*, *277*(5326), 684–686. https://doi.org/10.1126/science.277.5326.684

Kuhl, P. K., Tsao, F. M., & Liu, H. M. (2003). Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(15), 9096–9101. https://doi.org/10.1073/pnas.1532872100

ManyBabies Consortium. (2020). Quantifying Sources of Variability in Infancy Research Using the Infant-Directed-Speech Preference. *Advances in Methods and Practices in Psychological Science*, *3*(1), 24–52. https://doi.org/10.1177/2515245919900809

Margulis, E. H., Wong, P. C. M., Turnbull, C., Kubit, B. M., & McAuley, J. D. (2022). Narratives imagined in response to instrumental music reveal culture-bounded intersubjectivity. *Proceedings of the National Academy of Sciences*, *119*(4), e2110406119. https://doi.org/10.1073/pnas.2110406119

Mehr, S. A., Singh, M., Knox, D., Ketter, D. M., Pickens-Jones, D., Atwood, S., Lucas, C., Jacoby, N., Egner, A. A., Hopkins, E. J., Howard, R. M., Hartshorne, J. K., Jennings, M. V., Simson, J., Bainbridge, C. M., Pinker, S., O'Donnell, T. J., Krasnow, M. M., & Glowacki, L.

(2019). Universality and diversity in human song. *Science*, *366*(6468), eaax0868. https://doi.org/10.1126/science.aax0868

Mendoza, J. K., & Fausey, C. M. (2021). Everyday music in infancy. *Developmental Science*, *n/a*(n/a), e13122. https://doi.org/10.1111/desc.13122

Moon, C., Cooper, R. P., & Fifer, W. P. (1993). Two-day-olds prefer their native language. *Infant Behavior and Development*, *16*(4), 495–500.

Nakata, T., & Trehub, S. E. (2004). Infants' responsiveness to maternal speech and singing. *Infant Behavior and Development*, *27*(4), 455–464. https://doi.org/10.1016/j.infbeh.2004.03.002

Nakata, T., & Trehub, S. E. (2011). Expressive timing and dynamics in infant-directed and non-infant-directed singing. *Psychomusicology*, *21*(1–2), 45–53. https://doi.org/10.1037/h0094003

Nguyen, T., Flaten, E., Trainor, L. J., & Novembre, G. (2023). Early social communication through music: State of the art and future perspectives. *Developmental Cognitive Neuroscience*, *63*, 101279. https://doi.org/10.1016/j.dcn.2023.101279

Papadimitriou, A., Smyth, C., Politimou, N., Franco, F., & Stewart, L. (2021). The impact of the home musical environment on infants' language development. *Infant Behavior and Development*, *65*, 101651. https://doi.org/10.1016/j.infbeh.2021.101651

Partanen, E., Kujala, T., Tervaniemi, M., & Huotilainen, M. (2013). Prenatal Music Exposure Induces Long-Term Neural Effects. *PLOS ONE*, *8*(10), e78946. https://doi.org/10.1371/journal.pone.0078946

Politimou, N., Stewart, L., Müllensiefen, D., & Franco, F. (2018). Music@Home: A novel instrument to assess the home musical environment in the early years. *PLOS ONE*, *13*(4), e0193819. https://doi.org/10.1371/journal.pone.0193819

Rabinowitch, T.-C., & Meltzoff, A. N. (2017). Synchronized movement experience enhances peer cooperation in preschool children. *Journal of Experimental Child Psychology*, *160*, 21–32. https://doi.org/10.1016/j.jecp.2017.03.001

Ramirez-Esparza, N., Garcia-Sierra, A., & Kuhl, P. K. (2014). Look who's talking: Speech style and social context in language input to infants are linked to concurrent and future speech development. *Developmental Science*, *17*(6), 880–891. https://doi.org/10.1111/desc.12172

Richards, J. A., Gilkerson, J., Xu, D., & Topping, K. (2017). How Much Do Parents Think They Talk to Their Child? *Journal of Early Intervention*, *39*(3), 163–179. https://doi.org/10.1177/1053815117714567

Romeo, R. R., Leonard, J. A., Robinson, S. T., West, M. R., Mackey, A. P., Rowe, M. L., & Gabrieli, J. D. E. (2018). Beyond the 30-Million-Word Gap: Children's Conversational Exposure Is Associated With Language-Related Brain Function. *Psychological Science*, *29*(5), 700–710. https://doi.org/10.1177/0956797617742725

Romeo, R. R., Segaran, J., Leonard, J. A., Robinson, S. T., West, M. R., Mackey, A. P., Yendiki, A., Rowe, M. L., & Gabrieli, J. D. E. (2018). Language Exposure Relates to Structural Neural Connectivity in Childhood. *The Journal of Neuroscience*, *38*(36), 7870–7877. https://doi.org/10.1523/jneurosci.0484-18.2018

Rowe, M. L. (2012). A Longitudinal Investigation of the Role of Quantity and Quality of Child-Directed Speech in Vocabulary Development. *Child Development*, *83*(5), 1762–1774. https://doi.org/10.1111/j.1467-8624.2012.01805.x

RStudio Team. (2020). *RStudio: Integrated Development for R* [Computer software]. http://www.rstudio.com/.

Team, R. C. (2020). *R: A Language and environment for statistical computing.* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/.

Trainor, L. J. (1996). Infant preferences for infant-directed versus noninfant-directed playsongs and lullabies. *Infant Behavior and Development*, *19*(1), 83–92. https://doi.org/10.1016/S0163-6383(96)90046-6

Trainor, L. J., Clark, E. D., Huntley, A., & Adams, B. A. (1997). The acoustic basis of preferences for infant-directed singing. *Infant Behavior and Development*, *20*(3), 383–396. https://doi.org/10.1016/S0163-6383(97)90009-6

Tsang, C. D., Falk, S., & Hessel, A. (2017). Infants Prefer Infant-Directed Song Over Speech. *Child Development*, *88*(4), 1207–1215. https://www.jstor.org/stable/45046377

Weisleder, A., & Fernald, A. (2013). Talking to Children Matters:Early Language Experience Strengthens Processing and Builds Vocabulary. *Psychological Science*, *24*(11), 2143–2152. https://doi.org/10.1177/0956797613488145

Yan, R., Jessani, G., Spelke, E. S., de Villiers, P., de Villiers, J., & Mehr, S. A. (2021). Across demographics and recent history, most parents sing to their infants and toddlers daily. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1840), 20210089. https://doi.org/10.1098/rstb.2021.0089

Young, S. (2008). Lullaby light shows: Everyday musical experience among under-two-year-olds. *International Journal of Music Education*, *26*(1), 33–46. https://doi.org/10.1177/0255761407085648