RESEARCH ARTICLE

Comparison of speech and music input in North American infants’ home environment over the first 2 years of life

Lindsay Hippe¹,² | Victoria Hennessy¹ | Naja Ferjan Ramirez¹,³ | T. Christina Zhao¹,²

¹Institute for Learning and Brain Sciences, University of Washington, Seattle, Washington, USA
²Department of Speech and Hearing Sciences, University of Washington, Seattle, Washington, USA
³Department of Linguistics, University of Washington, Seattle, Washington, USA

Correspondence
T. Christina Zhao, University of Washington, BOX 357988, Seattle, WA 98195, USA.
Email: zhaotc@uw.edu

Funding information
National Institutes of Health, Grant/Award Number: R01DC020419

Abstract
Infants are immersed in a world of sounds from the moment their auditory system becomes functional, and experience with the auditory world shapes how their brain processes sounds in their environment. Across cultures, speech and music are two dominant auditory signals in infants’ daily lives. 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 of daylong audio recordings collected in English-learning infants’ home environments, at 6, 10, 14, 18, and 24 months of age. Using a crowdsourcing approach, 643 naïve listeners annotated 12,000 short snippets (10 s) 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. At every age point, infants were exposed to more music from an electronic device than an in-person source; this pattern was reversed for speech. The percentage of input intended for infants remained the same over time for music while that percentage significantly increased for speech. We propose possible explanations for the limited music input compared to speech input observed in the present (North American) dataset and discuss future directions. We also discuss the opportunities and caveats in using a crowdsourcing approach to analyze large audio datasets. A video abstract of this article can be viewed at https://youtu.be/lFj_sEaBMN4

KEYWORDS
auditory environment, crowdsourcing, infancy, LENA, music input, speech input

1 | 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 examine the auditory environment of infants.

Speech and music are two large components of the human soundscape, and universal across cultures (Hilton et al., 2022; Mehr et al., 2019). When interacting with infants, adults alter their communication to a special style commonly described as infant-directed (ID) speech and infant-directed 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). However, it is critical to note that there are many other dimensions to ID communication beyond acoustic modifications, such as simplified vocabulary and syntactic structure (Genovese et al., 2020) and a 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). When presented with audiovisual stimuli of ID song and speech, infants looked longer for ID singing than speech and more specifically, they exhibited more mouth-looking behavior for ID singing that became more prominent with age (Alviar et al., 2023; Nakata & Trehub, 2004). 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.

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 repeatedly demonstrated that both the quantity and quality of speech input in infancy, particularly ID speech, can have profound long-term impacts on infants’ language development (Cartmill et al., 2013; Rowe, 2012). Earlier studies used methodologies such as annotating short videos of mother-infant free play in the lab (Dave et al., 2018) or in home (Rowe, 2012). However, these methods were limited in the amount of data and contexts they could capture. In more recent years, researchers have further honed 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 gesture and language 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 music in infants’ environments (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 and 12 months. They found that more than half of everyday music for infants contains singing, and more than three-quarters contain instrumental music. 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 environments. Furthermore, it is unknown how speech versus music input may change over the course of infant development. The current study addresses this important gap by analyzing an existing LENA dataset collected longitudinally in North American, English-speaking families when infants were 6, 10, 14, 18, and 24 months. We describe speech and music input in parallel throughout infancy and address three main questions: first, we examine the quantity of speech input in comparison to music input over the first 2 years of life. Second, given the increasing prevalence of electronic devices in today’s world, we further examine the source of speech versus music input (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 measures of speech and music input, we opted to crowdsource LENA data annotation. This approach was 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 music in infants’ environments (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 and 12 months. They found that more than half of everyday music for infants contains singing, and more than three-quarters contain instrumental music. 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 environments. Furthermore, it is unknown how speech versus music input may change over the course of infant development. The current study addresses this important gap by analyzing an existing LENA dataset collected longitudinally in North American, English-speaking families when infants were 6, 10, 14, 18, and 24 months. We describe speech and music input in parallel throughout infancy and address three main questions: first, we examine the quantity of speech input in comparison to music input over the first 2 years of life. Second, given the increasing prevalence of electronic devices in today’s world, we further examine the source of speech versus music input (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 measures of speech and music input, we opted to crowdsource LENA data annotation. This approach was...
Methods

Experimental procedure

Participants

Stimulus

2 | METHODS

2.1 | 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 2 days to record their naturalistic sound environment for up to 16 hours per day at each recording age. They were recorded five times longitudinally at 6, 10, 14, 18, and 24 months of age, within 3 days of each target age on average. 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 subjects with complete sets of LENA recordings (i.e., N = 24 with 12 male infants, 2-day recordings at each of the 5 ages).

The LENA Advanced Data EXtractor (ADEX) tool was first used to divide each daylong recording into 5-min 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% of the 5-min segments ranked by the amount of Silence are predominantly silent (mean = 239.00, std = 88.15). Therefore, we dropped the top 10% of 5-min segments with the highest amount of Silence from each recording. From the remainder of each recording, we randomly selected 50 10-s snippets of audio, resulting in 100 10-s snippets per subject per age (see Figure S2 for an example of sampling). The length of the snippet (10 s) was chosen to mitigate the potential risk of including confidential information, while still having enough information for participants to annotate. In total, 12,000 snippets across 5 ages (i.e., 2,400 snippets per age) were used as the stimuli for the current annotation experiment.

2.2 | 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 credit 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.

2.3 | Experimental procedure

The current annotation experiment was conducted on the Zooniverse, an online platform specialized for 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 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 given the choice to listen to either 50 or 100 snippets to receive 0.5 or 1 course credit through the ORPP system. For each snippet, they answered a series of questions (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. An individual snippet ceased to be presented to participants (i.e., was “retired”) once three different participants selected the None/Other option OR once the snippet received five votes from different participants.

Two levels of randomization were used for snippet 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 2 years (“Speech” from this point), (2) Music or singing (“Music” from
this point), (3) Speech produced by someone older than 2 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 fourth options (i.e., None). Participants continued to answer the other two main questions if they chose options 1–3. Further, if they chose the third option (Speech AND Music), they answered the following questions twice, first for speech and then 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.

2.4 Data processing

Raw annotation data (.csv files) were exported from Zooniverse and processed using in-house Python scripts (Hippe and Zhao, 2023, Tots & Tunes: https://osf.io/84rt9). Specifically, we first transformed the data structure such that each row contained the aggregated annotation data for each snippet from all participants who made annotations for that snippet. The percentage of votes for each option was calculated and stored for each question and for each snippet, given the variation across snippets in terms of the number of total votes received. 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 is 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% of the vote and, if so, that option was marked 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 [20, 20, 0, 60], no option has over 50% 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% of the vote, the maximum percentage. To address this, the value for the third 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 (first pass overall: 90.2%, second pass overall: 2.2%, third pass overall: 4%) with a small percentage unresolved (unresolved overall: 3.2%, see Table S1 by age group).

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 versus 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 (number of “None” snippets for each age: 6 months: 924, 10 months: 958, 14 months: 882, 18 months: 830, and 24 months: 782, see Figure S3 for the distribution of different categories by age). Only snippets with a final annotation of Speech, Music, or Speech AND Music were entered into further majority voting to derive the 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 (Response options: In person, Electronic device, Both). 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 third option and either of the first two options, the same strategy used for the first question was employed by adding the number for the third 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 (Response options: Yes, infant-directed vs. No, not infant-directed), 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/Singing) 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/Singing) to the other two categories (e.g., Speech category and Music/Singing category). For example, if one infant at 6 months received a 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 second and third 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,” so 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/singing, (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/singing, (8) music/singing through an electronic device, (9) music/singing directed to a baby, and (10) music/singing not directed to baby.

2.5 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 (Note that all comparisons are within-subject in nature). 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 versus 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. Post hoc pair-wise comparisons were conducted for fixed factors. Lastly, exploratory correlation analyses were conducted between total speech and total music across participants at each age.

3 RESULTS

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

The amount of total speech versus 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) ($F = 2106.41, p < 0.001, \beta = 35.12, \eta^2 = 0.98$) and a significant interaction between Age and Type of input ($F = 5.88, p = 0.002, \beta_s = 6.54, 6.71, 9.13, 13.83, \eta^2 = 0.01$). 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, $p = 0.03$).

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 versus 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) ($F = 1435.69, p < 0.001, \beta = -3.33, \eta^2 = 0.35$), a significant effect of Input Source (In Person vs. Electronic Device) ($F = 955.10, p < 0.001, \beta = -6.79, \eta^2 = 0.23$), a significant interaction between Age and Type of input ($F = 4.07, p = 0.003, \beta_s = 1.58, 2.0, 2.23, 1.38, \eta^2 = 0.004$), a significant interaction between Age and Input Source ($F = 3.64, p = 0.006, \beta_s = 0.88, 1.0, 1.75, 0.96, \eta^2 = 0.004$), a significant interaction between Input Type and Source ($F = 1694.56, p < 0.001, \beta = 42.25, \eta^2 = 0.41$) and critically, a significant 3-way interaction ($F = 2.71, p = 0.03, \beta_s = 4.46, 3.63, 4.75, 11.63, \eta^2 = 0.003$). 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 ($p > 0.1$). For speech, input from Electronic Device sources remained stable across ages ($ps > 0.1$) while input from In-Person Sources increased (6–10 mo: $p = 0.025$, 10–24 mo: $p < 0.001$, 14–24 mo: $p < 0.001$, 18–24 mo: $p = 0.003$).

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 versus 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) ($F = 1164.56, p < 0.001, \beta = 11.92, \eta^2 = 0.89$), a significant effect of Input Recipient (Infant vs. Not Infant) ($F = 46.27, p < 0.001, \beta = 3.71, \eta^2 = 0.04$), a significant interaction between Age and Type of input ($F = 3.65, p = 0.006, \beta_s = 4.83, 6.50, 9.21, 17.63, \eta^2 = 0.01$), a significant interaction between Age and Input Recipient ($F = 10.40, p < 0.001, \beta_s = -0.79, 0.79, 0.42, -0.38, \eta^2 = 0.03$), and critically, a significant 3-way interaction ($F = 9.96, p < 0.001, \beta_s = -0.387, -0.721, -10.625, -21.46, \eta^2 = 0.03$). 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 were intended for the infant ($ps > 0.1$). 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: $p < 0.001$).

4. Exploratory question: No significant correlation between speech and music input across infants

The Pearson correlation coefficients 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.

4.1 DISCUSSION

The present study quantitatively assessed speech and music input in North American infants’ naturalistic auditory environment over the first 2 years of their lives. By utilizing a crowdsourcing approach, short snippets were randomly sampled from daylong LENA recordings and annotated by multiple naive listeners with final annotations derived through a majority voting procedure. Our results addressed three important research questions: (1) How does the quantity of speech input compare to music input across the first 2 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 vs. through an electronic device) revealed significantly more in-person than electronic speech and significantly more electronic than in-person music. Further, in-person speech increased over time while all other 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 crossover 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 consistently less than non-ID music.

All of our findings regarding speech input are well aligned with the existing literature (e.g., Bergelson et al., 2023). The increasing amount of speech infants hear with development highlights its dominance in infants’ auditory environments. 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 by Bergelson and colleagues (2019), where they also demonstrated a decrease of non-ID speech as infants age. However, in their study, the IDS amount remained stable with age while the IDS amount in our study increased significantly. Further, in our data, the crossover (i.e., the quantity of
ID speech surpasses the quantity of non-ID speech heard by infants) was observed around 18 months of age. This coincides with the drastic increase of conversation turn-taking known to spike at 18 months, and supports the idea that as infants become more communicative, their caregivers may naturally start engaging with them in more verbal and social communication, which may further propel infants’ language acquisition (Ferjan Ramirez et al., 2021). Given the importance of social, in-person, and infant-directed speech input (Ferjan Ramirez et al., 2020; Kuhl et al., 2003), it is reassuring to see that parents are indeed highly engaged as infants’ language skills improve. Development in language skills could also extend to the development of music skills (Liu et al., 2023), making it an interesting future direction.

However, our findings on music input, especially when compared to speech, are surprising. Overall, we found that infants in North American, English-speaking families engage in very few live musical activities with their caregivers that are specifically intended for the infant. First, the low occurrence of music was unexpected, given previous research in which parents report singing or playing recorded music on a daily basis to their infants (Yan et al., 2021), and existing studies that document infant-directed singing as a cross-cultural phenomenon (Hilton et al., 2022; Mehr et al., 2019). Our results provide more fine-grained information on daily music input and suggest that while these behaviors exist within most families (e.g., the number of infants with 0 snippets coded as infant-directed music at each age is very low (6 mo.: 3, 10 mo.: 3, 14 mo.: 1, 18 mo.: 1, 24 mo.: 3), they occur at a very low frequency. Further, it is also interesting to observe more snippets labeled as “Music through an electronic device” than “In-person music or singing,” consistently across ages. This is a little different from Yan et al. (2021) where parents reported an equal likelihood of singing live or playing recorded music to their infants, with a large decrease in live singing as infants get older. Future research is needed to understand the various contexts (e.g., naptime vs. playtime) in which different sources of music (e.g., live singing vs. music played through a device) appear in infants’ daily lives. Future works should also explore whether and how infants respond differently to live versus recorded music in their natural auditory environments. Even though infants have been reported to learn music through passive exposure (i.e., playing music in the background) (Hannon & Trehub, 2005), it has also been shown that active music activities (i.e., interactive musical play with infants) are more effective for learning (Gerry et al., 2012).

Finally, a lack of correlation between speech and music input suggests that music may be contributing to infant development independently from speech. Future research should consider both speech and music input when studying infants’ auditory environment. It is possible that all results observed in this study are specific to this sample, given the participating families were homogenous both geographically and culturally. 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.

Given the limited amount of music input observed in the current study, our findings point to significant need and potential for improving music input in North American infants’ auditory environments, especially in the form of live, interactive, and infant-directed musical experiences. Over the past decade, a growing and converging body of evidence has documented the multifaceted benefits of active music experience for other aspects of early development, such as promoting speech and language learning (Zhao et al., 2022; Zhao & Kuhl, 2016) and social emotional connections between infants and their caregivers and peers (Cirelli et al., 2014; Rabinowitch & Meltzoff, 2017). Further studies might first seek a deeper understanding of the reasons behind the general paucity of music in North American, English-speaking families with young infants. Such insights will be necessary to design parent-focused intervention methods to increase infant-directed musical activities and maximize the benefits of early music experience.

Methodologically, we took a new approach to annotate a large corpus of daylong LENA audio recordings via crowdsourcing. Traditionally, this type of work is done in-lab and by a group of trained research assistants, which is a reliable, but labor and cost-intensive practice. To ensure accuracy from a large group of naïve annotators, we implemented a majority voting procedure. Indeed, a reliability analysis validated the annotations derived from the current approach given that they were highly consistent with annotations from a trained coder. Thus, we are confident in the potential of crowdsourced annotations for analyzing LENA recordings in future work. However, given that it is a relatively new approach to annotate audio data on Zooniverse (Semenzini et al., 2021), we also acknowledge several important caveats that we learned over the course of the study and hope future studies can improve upon. First, we took a relatively small, random sample across 90% of the recordings (totaling 33.3 h of audio for 24 infants at 5 ages), with the goal of extracting dependent measures to quantitatively compare speech and music input. While similar sparse sampling approaches are commonly used in LENA data analyses (Bergelson et al., 2019; Cychosz et al., 2021; Ferjan Ramirez et al., 2020; Ramirez-Esparza et al., 2014), our sample may not be completely representative of the entire daylong recording and could be misleading if used to extrapolate the total duration of speech and/or music experience (see Figure S4 for a simulation experiment demonstrating the potential underestimation of total duration when extrapolating from the samples). Another reason the present study does not attempt to estimate total duration of speech or music is the binary nature (e.g., speech vs. no speech) of our annotations because marking that a snippet contains speech does not capture the length of the speech occurrence (e.g., 1 s. vs. 10 s.). Indeed, Bergelson et al. (2019) discussed similar issues when annotating ~4 h of data out of 61 recordings. To accurately estimate duration or percentage of the day for a certain type of auditory input, future studies will need to implement a more fine-grained annotation strategy utilizing a much higher percentage of the entire audio dataset. For example, Mendoza and Fausey (2021) manually annotated 467 h of recordings across 35 infants to confidently estimate duration of daily music experience. Utilizing a significantly larger percentage of samples could be more feasible with improved automated algorithms to flexibly remove silence and noise segments. In our study, close to half of the snippets were identified as “None” and were excluded from further annotations. Reliable and accurate automatic removal of those segments would not only help target our resources to annotate the important segments in...
more detail, but would also improve duration estimate through extrapolation (see Figure S4, in which a simulation experiment demonstrates that when "Silence" data are partially removed before sampling, the estimated duration is much closer to ground truth).

A related caveat is that music is likely slightly underrepresented in the current study due to several factors. First, post hoc simulation experiments (Figure S4 and Supplemental Method) show that while our sampling method robustly captures the quantitative difference between speech and music, music is slightly more underrepresented than speech. This might be because when there is overall much less music and our sampling method is sparse, there is a higher chance for music to be left out of the randomly sampled snippets. Second, when the "Both" option is selected in the first question (Figure S3), annotators were always asked to evaluate the speech content before the music content in subsequent questions, which may have resulted in less precise labeling of the music co-occurring in the same snippet.

Second, even though we used only a few and relatively simple response categories (e.g., Speech, Music, None); ambiguity still existed. Examples of ambiguity included background speech that was unintelligible, speech directed to another child, or speech delivered in a "rap-like" manner. Our trained coder has compiled such instances (~2% of the dataset), which will help us to further improve our annotation instructions. Critically, the majority of the ambiguity concerns the line between speech and music. It can be particularly blurry in infant-directed vocalizations, especially when trying to decipher "sing-songy" speech or when caregivers speak in a highly rhythmic manner (e.g., reading a nursery rhyme). To better understand where the line is across listeners, targeted experiments are needed in which inputs are systematically manipulated (e.g., pitch variability) and classifications can be examined as a function of the input characteristics.

Third, the naive listeners in this study were asked to judge whether the speech and/or music was directed to an infant (i.e., intended target); however, 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 with full certainty whether the input was indeed intended for the infant without more information (e.g., visual cues or context from the caregiver). For example, parents might adopt an infant-directed speech register when talking to another child or a pet, which could be easily categorized as infant-directed. Without knowing the ground truth from the original speaker, strong agreement among coders does not necessarily mean correct judgment. This issue is not specific to this study but may apply to any annotation scheme that uses LENA recordings without direct context from caregivers. It is worth noting that in Bergelson et al. (2019), all annotators were research assistants highly trained in speech analysis and were asked to judge whether the speech was in a "Child-Directed Speech register (although they were permitted to use context, too)" instead of "Is it (the speech/music) directed to a baby?". It is possible that the variation in the annotation instruction (i.e., coding for speech register vs. intended target) led to differences in our findings regarding IDS amount across infant ages. Future systematic experiments are warranted to clearly understand the potential effects of annotation instruction. On the other hand, annotating intended recipients for music, particularly instrumental music, can be much harder. It is unclear what criteria were used by coders in our current annotation experiment. 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 would not capture the musical routines or preferences of a specific household (for instance, using Taylor Swift songs for play or a daily routine) or their culture (e.g., children's songs in other cultures), resulting in inaccurate/biased labeling. Several important questions should be addressed in future studies, such as asking follow-up questions regarding the reasoning behind an "infant-directed" labeling (e.g., acoustic cues, context, others), examining the individual variability across caregivers in their infant-directed vocalizations, and conducting initial surveys to assess what annotators/caregivers consider to be infant-directed music.

In conclusion, the current study provided a first look at infants’ speech and music input in their naturalistic home environments and observed large differences between the two domains. We believe these results demonstrate the need to further examine music input among infants across cultures in relation to their speech input; and provide motivation for finding ways to enrich auditory experiences in early development.

AUTHOR CONTRIBUTION
Conceptualization: T. Christina Zhao; Data curation: Naja Ferjan Ramirez; Formal analysis: Lindsay Hippe; Methodology: Lindsay Hippe, Victoria Hennessy, T. Christina Zhao; Software: Victoria Hennessy; Visualization: Lindsay Hippe, Victoria Hennessy; Writing–original draft: Lindsay Hippe, T. Christina Zhao; Writing–review & editing: Lindsay Hippe, Victoria Hennessy, T. Christina Zhao, Naja Ferjan Ramirez.

ACKNOWLEDGMENTS
Research reported in this publication was supported by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health (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 www.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.

CONFLICT OF INTEREST STATEMENT
The authors report no conflict of interest.

DATA AVAILABILITY STATEMENT


SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Hippe, L., Hennessy, V., Ramirez, N. F., & Zhao, T. C. (2024). Comparison of speech and music input in North American infants’ home environment over the first 2 years of life. Developmental Science, e13528. https://doi.org/10.1111/desc.13528