The Acoustic Characteristics of Infant-Directed versus Adult-Directed Speech and Song

Lindsay Hippe

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2026

Committee:

Christina Zhao

Amy Pace

Richard Wright

Program Authorized to Offer Degree:

Speech & Hearing Sciences

University of Washington

Abstract

The Acoustic Characteristics of Infant-Directed versus Adult-Directed Speech and Song

Lindsay Hippe

Chair of the Supervisory Committee:

Christina Zhao

Department of Speech & Hearing Sciences

Infant-directed (ID) speech is characterized by increased pitch, greater pitch variability, slower rate, and greater vowel space compared to adult-directed (AD) speech. Less is known about ID song, but research suggests that it exhibits increased mean pitch and longer vowel durations when compared to ID speech. The present study investigates the acoustic characteristics of the corner vowels (/i/, /u/, and /ɑ/) in ID and AD song and speech. ID vocalizations were characterized by significantly higher fundamental frequency (f0), duration, vowel space, global vowel dispersion, and within-category vowel dispersion compared to AD vocalizations. Similarly, song was associated with significantly higher measures of f0 and length than speech but was uniquely characterized by reduced f0 variability and increased duration variability when compared to speech. The increase in f0 and f0 variability associated with infant-directedness was much smaller in song than in speech, suggesting that acoustic modifications associated with ID vocalizations are constrained in song.

Dedicated to Bryce Erica Hippe

It is well established in the literature that, across cultures, caregivers modulate their speech when addressing infants (Fernald et al., 1989, Greiser & Kuhl, 1988, Kuhl et al., 1997, Kitamura et al., 2001, Cox et al., 2023, Hilton et al., 2022). This unique register is termed "parentese" or infant-directed speech (ID speech). A recent review of the literature concluded that ID speech is characterized by higher pitch, greater pitch variability, slower rate, and greater overall vowel space (Cox et al., 2023) compared to speech directed to adults. These acoustic differences have been shown to facilitate vocabulary development (Thiessen, Hill, & Saffran, 2005, Ferjan Ramírez, Lytle, & Kuhl, 2020) and to be preferred by infants cross-culturally (Werker et al., 1994, Byers-Heinlein et al., 2021).

Speech is not the only way that caregivers communicate with their infants: infant-directed song (ID song) is also an important component of infant-directed communication. However, there is much less research analyzing the acoustics of infant-directed song. Interestingly, many authors have characterized ID speech as song-like (Alviar et al., 2022, Hilton et al., 2022) due to its slower tempo, exaggerated pitch, and increased repetitiveness/rhythmicity. While there are certainly similarities between ID speech and ID song, namely a slower tempo compared to adult-directed versions (Tsang, Falk, & Hessel, 2016), the research that does exist, detailed in the following sections, has demonstrated differences between ID speech and ID song, as well as adult-directed speech and song (AD speech and AD song). Nevertheless, the evidence is still inconclusive, and further research is needed.

Although previous research has described the suprasegmental characteristics of ID and AD speech and song, less is known about their segmental characteristics. Both ID and sung vocalizations are known to be associated with changes in the vowel space (Cox et al., 2023; Bradley, 2018), and there is debate regarding whether ID vocalizations are produced with more

distinct vowel categories or more exaggerated, overlapping vowel categories when compared to AD vocalizations. (Martin et al., 2015). Analyzing vowels allows for investigation of the three main acoustic correlates of ID vocalizations (increased pitch, duration, and vowel space) at the segmental level. Therefore, the present study examines the acoustic characteristics of the corner vowels, which form the boundaries of the vowel space. The following sections review the extant literature on vowel acoustics in ID and AD speech and song.

Infant-Directed Speech and Song versus Adult-Directed Speech and Song

Falk & Audibert (2021) aimed to unveil the "acoustic signatures" (combinations of acoustic characteristics) that were associated with different types of everyday interactions between infants and their mothers. In this study, the authors focused on the acoustic characteristics of vowels in infant-directed versus adult-directed input across different types of interactions (arousing versus calm) and different vocalization types (song versus speech). They analyzed a corpus of recordings of 15 German-speaking mothers performing two types of spoken input (a rhyme, a read story) and two types of songs (a playsong, a lullaby) for their six-month-old infants (infant-directed condition) or for an experimenter (adult-directed condition). The analysis revealed a clear acoustic signature for interactive stimulation (arousing versus calm) in that both fundamental frequency (f0) and its variability were significantly higher in arousing compared to calm interactions. They also observed a smaller effect of vowel durational variability, with higher variability in calm compared to arousing interactions. A clear acoustic signature was also observed in vocalization type (song versus speech), with vowels being significantly longer in song compared to speech. Additional acoustic qualities characterized this dimension, with f0, f0 variability, duration variability, and vowel variability all being higher in speech compared to song. However, these effects were smaller than the effect of vowel duration.

Notably, there was not a clear set of acoustic qualities that differed significantly between infant-directed compared to adult-directed input. All acoustic variables except for vowel clarity varied as a function of infant-directedness, but the effect sizes were observed to be modest to small compared to the effects observed in the other dimensions (vocalization type versus interactive stimulation). Overall, the findings of this experiment are consistent with those of Falk et al. 2021, further highlighting the unique acoustic qualities of infant-directed speech and song. The authors suggest that these acoustic signatures may help infants to learn about paralinguistic functions and differentiate between spoken and musical registers.

Additional analysis of this same corpus has provided insight regarding the vowel space in ID and AD singing and speech. Audibert & Falk (2018) plotted the vowel space of the three corner vowels in the corpus and found higher within-vowel variability in both ID speech and ID song when compared to AD speech and song. They also noted that F1 was expanded while F2 was compressed in the ID versions compared to the AD versions of the stimuli. Finally, the most significant difference reported between the ID and AD stimuli was a higher fundamental frequency (f0) in ID compared to AD stimuli, which makes sense given that ID speech is known to be characterized by higher pitch compared to AD speech. Similarly, the authors describe how ID input is overall more acoustically variable than AD input, which is also in line with the consensus that ID speech is more variable than AD speech. Finally, the authors describe certain changes when comparing speech to song. First, f0 was higher in the playsong when compared to the story reading, which the authors suggest may be due to the specific melodic structure of the song and the content of the story. Second, the authors reported that within-vowel-category dispersion was higher in speech than in song, whereas vowel space expansion was overall smaller in song than in speech. The authors put forward a few possible explanations for this

phenomenon and highlight the need for further research to investigate the differences in more detail.

Trainor et al. (1997) aimed to uncover the acoustic characteristics that cause infants to prefer infant-directed singing. To do this, they recorded 15 mothers singing songs of their choice in both the presence and the absence of their infants and performed acoustic analyses of the resulting recordings. By analyzing the vowels in the recordings, they found that the mean pitch (Hertz [Hz]), jitter (fundamental frequency variation), and shimmer (amplitude perturbation), were higher in the infant-directed compared to infant-absent versions. Overall, ID vocalizations demonstrated higher pitch and greater variation in both fundamental frequency and amplitude, which is also in line with patterns described in the research concerning AD versus ID speech. Overall, given the consistencies between these studies, it appears that there is an interaction between infant-directedness and vocalization type, which future studies may tease apart more thoroughly.

Infant-Directed Speech versus Infant-Directed Song

Tsang, Falk, & Hessel (2016) compared ID speech and ID song in both English and Russian. The authors used a corpus of recordings from 10 Russian-speaking mothers and a separate corpus from eight English-speaking mothers that contained samples from play songs, lullabies, rhymes, and conversations of them singing and speaking to their infants during their first year. For the Russian corpus, native English-speaking adults were asked to rate 49 samples on a 7-point Likert scale (1 = definitely speech and 7 = definitely singing), and the samples that were most definitely speech or singing were chosen as the stimulus for analysis. As a result of this process, six samples were chosen for ID speech (mean duration = 2.06 s, s =0.87) and five samples were chosen for ID song (mean duration = 2.93 s, s = 0.62). The same number of

samples were taken for each vocalization type (ID speech/ID song) from the English corpus of 41 total samples, and they were chosen because they most closely matched the acoustic differences between ID speech and ID song observed in the Russian corpus. The authors did not report the length of the English samples. The measures pertaining to vowels that were analyzed were vowel pitch variability (semitones), and the overall percent of vowels. Overall, the authors found that vowel pitch variability was significantly higher in ID speech than ID song, but the percent of vowels was higher in ID song than in ID speech.

There are other studies that examined acoustic features in the context of infant perception and attention. Falk et al. (2021) examined the acoustics of infant-directed song versus speech in the context of processing speech information. A female native speaker of German was recorded in a soundproof booth producing sequences of two syllables in both ID speech and ID song. For the purposes of the experiment, one recording of repeated syllables and one recording of alternating syllables were presented for each condition (ID speech/ID song). Though the study itself did not aim to examine the acoustic differences between the conditions, the authors reported that the vowel duration variability was notably higher in sung versus spoken stimuli.

Some studies have examined these differences in the context of infant attention. Alviar et al. (2022) investigated how ID song versus ID speech potentiate infant attention to adults' mouths. Their stimuli consisted of 16 total audiovisual clips, each showing one of five different female actors looking directly into a camera and producing either ID speech or ID song. The authors measured rhythmic variability using the Normalized Pairwise Variability Index (nPVI), which "quantifies the average difference between adjacent events in a sequence (e.g. musical notes in a melody, successive vowels in a spoken sentence)" (Daniele, 2017). The authors

reported that the rhythmic variability (nPVI) of vowels was higher in ID speech than ID song, corroborating the results reported by Falk et al. (2021).

Overall, there is a lack of research focusing on vowels in infant-directed song and speech. However, the studies that do exist indicate that that there is higher variability in both pitch and timing in ID speech compared to ID song, patterns can be seen across these studies, there is no clear agreement. The sparse nature of the extant literature highlights the need for additional research that systematically compares ID speech and ID song to clarify how these two registers differ acoustically in early caregiver–infant interaction.

The Present Study

The present study aims to address limitations of previous work and further clarify how acoustic changes associated with infant-directedness operate in song. Many of the studies that analyze the acoustics of infant-directed song and infant-directed speech rely on limited stimulus materials, with many studies incorporating only one or two songs, which limits the generalizability of their findings. Additionally, since the stories and songs used in these studies were largely intended for children, adult-directed vocalizations may still exhibit acoustic characteristics associated with infant-directedness. The present study addresses this issue by including six different songs as well as both scripted speech and prompts for naturalistic speech that could reasonably be directed to adults as well as infants. Another issue with many of the studies presented is that an infant's presence during the production of ID speech and ID song may affect the acoustic signal, as research has demonstrated how caregivers adjust their behavior in turn with their infant's behavior (Gros-Louis et al., 2006, van der Klis, Adriaans, & Kager, 2023). This study controlled for the possibility of an infant's emotional state affecting their mothers' vocalizations by presenting caregivers with images of their infant while smiling.

Present Research Questions

The present study asks three questions that aim to describe the acoustic characteristics of the corner vowels (/i/, /u/, and /ɑ). Corner vowels were chosen because they occur in many of the world's languages (Liljencrants & Lindblom, 1972) and form the boundaries of a triangle approximating the vowel space (Kuhl et al., 1997).

1. How do (1) vowel fundamental frequency (f0), (2) vowel f0 variability, (3) vowel duration, (4) vowel duration variability, (5) global vowel dispersion, (6) within-category vowel dispersion, and (7) vowel space area differ between AD versus ID vocalizations across both speech and song?
2. How do these same variables differ between speech and song across both AD and ID vocalizations?
3. Is there an interaction effect between vocalization target (AD versus ID) and vocalization type (speech versus song) on each of the dependent variables?

Hypotheses

Question 1 (AD versus ID). ID vocalizations will have (1) higher f0, (2) greater f0 variability, (3) longer vowel duration, (4) greater vowel duration variability, (5) greater global vowel dispersion, (6) greater within-category vowel dispersion, and (7) greater vowel space area compared to AD vocalizations.

Question 2 (Speech versus Song). Song will have (1) higher f0, (2) lower f0 variability, (3) greater vowel duration, (4) lower vowel duration variability, (5) lower global vowel dispersion, (6) lower token-level vowel dispersion, and (7) lower vowel space area compared to speech.

Question 3 (Target x Type Interaction). There will be an interaction effect between vocalization target (AD versus ID) and vocalization type (speech versus song) on each of the dependent variables.

Methods

Participants

The participants were 15 infants between 5-6 months of age and their mothers. They were recruited primarily through emails sent to families enrolled in the University of Washington's Communication Studies Participant Pool and secondarily through posters distributed to various online groups, local parent groups, and email lists. All recruitment materials described the study as an infant-caregiver interaction study, with the goal of the study described as "to learn how parents interact with their infants and potentially inform early language learning strategies." The study protocol did not mention anything about speech or music to avoid influencing caregiver behavior.

Inclusion criteria were as follows: 1) the infant was the first born with no siblings, 2) born within two weeks of their due date with a normal birth weight of 6-10 pounds, 3) from a predominantly monolingual English-speaking family, 4) typically developing, and 5) with no major health concerns, or 6) known major cognitive or neurodevelopmental disorders in their immediate family. Once the participants completed their appointments, a small bath toy was given as a prize to the infant, and the caregivers were compensated $40 each through Tango gift cards for their participation.

Materials

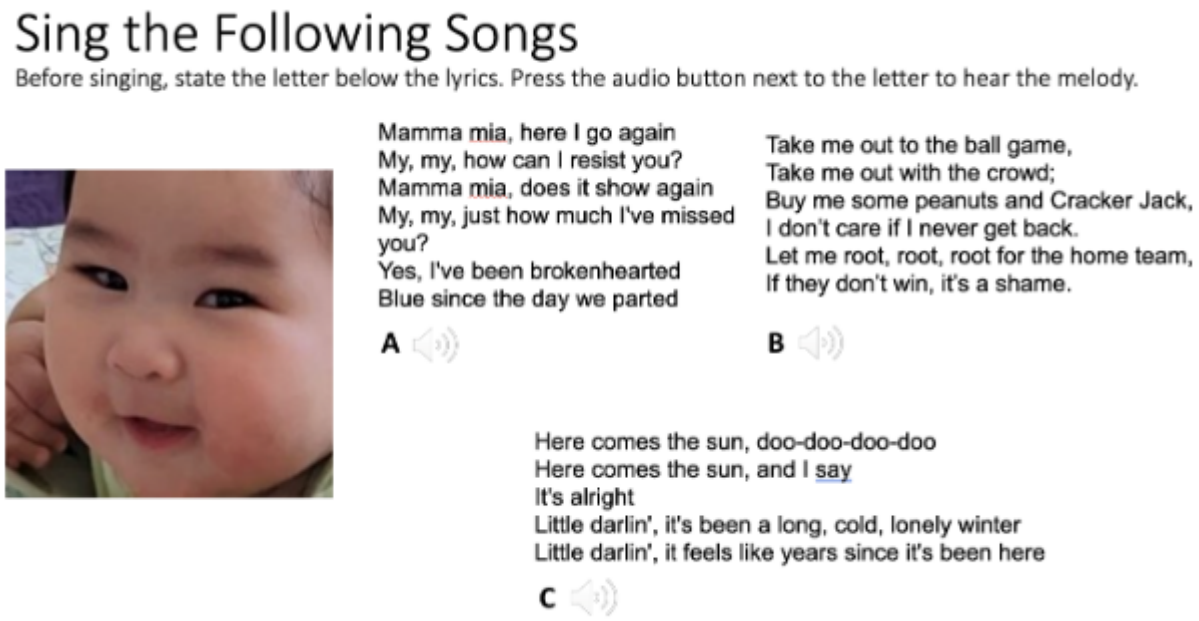
All experimental tasks were completed in a 9 x 10-foot sound attenuating booth. The whole experiment consisted of three tasks: 1) free play, 2) simulation, and 3) musical aptitude

test. Only the materials for the simulation task will be described here. Four HD video cameras (Marshall CV505/6, sampling rate of 29.97 FPS) at different angles recorded video, and an omnidirectional ceiling-mounted microphone connected directly to the cameras recorded audio to synchronize all audio recordings to video streams. Caregiver speech was recorded using a DPA 4060 omnidirectional lapel microphone connected via a microdot XLR adapter to a Zoom H4n Pro Recorder running phantom power. All audio data was digitized at 44.1 kHz. Sony MDR-7506 headphones were used to present audio to the participants, and a 9th Generation iPad running on iOS 17.3 was used to present the visual stimuli. The visual stimuli consisted of a slide deck of 13 slides that were prepared before the appointments and presented to the participants on the iPad. The first 12 contained caregiver instructions, while the final slide indicated that the task was complete, and they could exit the booth.

The simulation task included three subtasks: 1) speech, 2) singing, and 3) clapping. Each subtask consisted of four slides: two infant-directed and two adult-directed. Each pair of slides within a subtask was identical in content and format; the only difference was the image displayed to indicate the intended audience. These images, provided by the caregiver beforehand, depicted either their own infant or an adult loved one. Each image was cropped to a square format with the face centered and added to the slide. Slide layout was kept consistent across all participants, and the order of slides was randomized for each session.

Figure 1:

Example Slide from Slide Deck Provided to Participants



Singing Subtask. Caregivers were presented with the lyrics of the choruses from six songs: A) "Mamma Mia," B) "Take Me Out to the Ball Game," C) "Here Comes the Sun," D) "Hallelujah," E) "Dancing Queen," and F) "You've Got a Friend in Me." These songs were judged by the research team to be easily recognizable and could reasonably be directed to either an adult or an infant. They also spanned a range of time signatures: 3/4 (Take Me Out to the Ball Game), 6/8 (Hallelujah), and 4/4 (all others). The corner vowels /i/, /u/, and /ɑ/ were balanced across the songs as well. Across all six songs, there were 20 instances of the vowel /i/, 18 instances of /u/, and 19 instances of /ɑ/. When counting only vowels occurring on strong beats (beats 1 and 3), the totals were 18 for /i/, 14 for /u/, and 15 for /ɑ/. Each slide in the singing subtask presented three songs: one slide featured songs ABC, and another featured songs DEF. Each song was accompanied by a speakerphone icon that would play the piano version of the

melody of the choruses. These piano recordings were composed and produced by the research team using Musescore Studio (Version 4.4.4; Schweer, 2024).

Clapping Subtask. Caregivers were presented with the same songs that were used for the singing task and instructed to clap the rhythm of each song. This part of the data will not be analyzed in the present study.

Speech Subtask. For the speech subtask, there were two prompts: 1) scripted speech and 2) naturalistic speech. The prompt for the scripted speech task was the same for all caregivers and was written such that each of the corner vowels was included 17 times. For the naturalistic speech task, the participant was prompted to respond to one of the following prompts: 1) talk about your weekend, 2) talk about people important to you, or 3) talk about the time of day.

Procedures

Potential participants were invited to complete a screening form online. Families who met the inclusion criteria were contacted via phone by a lab team member to discuss the study and schedule an appointment. Prior to coming to the lab for appointments, participants completed both an informed consent and a demographic questionnaire form. In the consent form, participants uploaded images of both their infant and an adult loved one with whom they have a close relationship. They were instructed to choose images in which their loved ones were happy and to ensure that their faces were clearly visible. In the demographic information form, participants answered questions about their infants. The questions involved a variety of topics (health, interactions, childcare, language, music, digital media, sleep, nutrition, and mindfulness) in order to obscure the specific research focus on music.

When participants arrived for their appointments, they were shown the space and informed that they could pause the session at any time (e.g., feedings or diaper changes) The experimenters then gave a brief overview of the experimental tasks and placed the lapel

microphone on the caregiver's shirt. While the full protocol included three tasks (free play, simulation, and musical aptitude test), only the procedures for the simulation task will be presented here. Caregivers completed this task by responding to the slide prompts as if either their infant or adult loved one were physically present. They wore the lapel microphone and used headphones to hear the piano melodies. For unfamiliar songs, caregivers were encouraged to press the speakerphone icon to hear the melody, then attempt the song to the best of their ability. The clapping slides included only rhythmic clapping instructions. This task typically took about 15-20 minutes, and the whole session typically took about an hour and a half.

Once the simulation task was completed, the microphone was removed, and the cameras were turned off. Caregivers then completed a short task in which they were asked to rate their comfort level with the singing components of the simulation task on a 5-point Likert scale (1 = least comfortable, 5 = most comfortable).

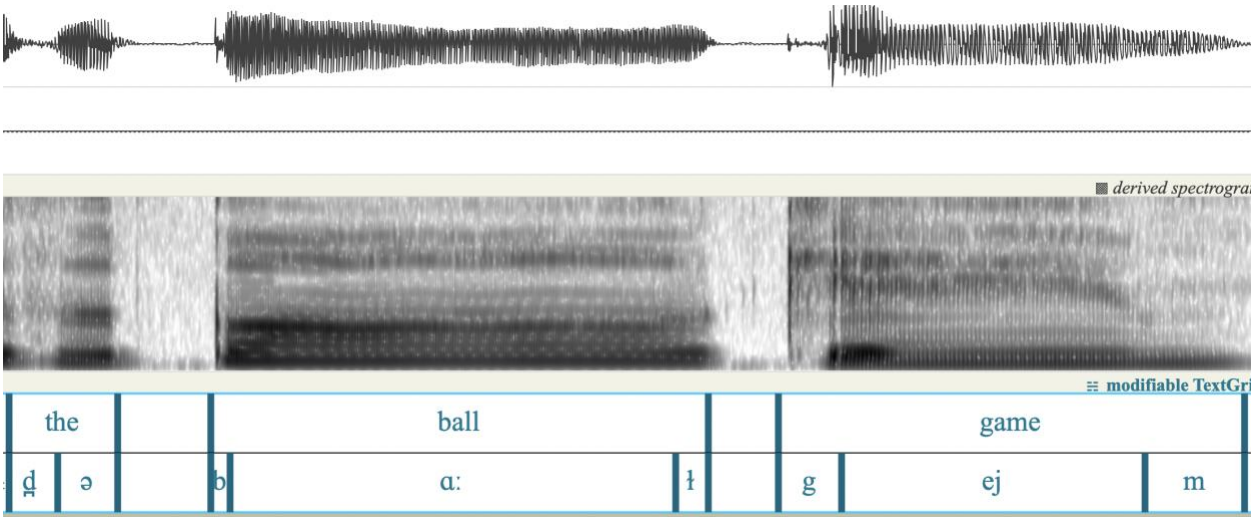
Data Analysis

All audio files, except for naturalistic speech, were automatically transcribed using WhisperX (Bain et al., 2023) and then checked by trained undergraduate research assistants (RAs). Naturalistic speech audio files were transcribed by a trained undergraduate TA. Vowels were extracted using the Montreal Forced Aligner (MFA; McAuliffe et al., 2026). The US English pronunciation dictionary and the English acoustic model were used to generate the MFA output, which consisted of textgrid files containing two tiers: one with words and one with phone-level alignments. Manual adjustments were made to the alignments generated by the MFA in Praat (Boersma & Weenink, 2025) by the author and trained RAs. Prior to acoustic extraction, trained undergraduate RAs reviewed and corrected the phoneme boundaries generated by the

Montreal Forced Aligner (MFA) in Praat to ensure that each interval contained only the target vowel. Figure 2 visualizes a phonetically aligned song excerpt in Praat.

Figure 2

Phonetically Aligned Phrase



Note: The first row is the word tier, and the row below is the phone-level alignment after it was manually adjusted.

During this adjustment process, vowels produced with creaky or breathy phonation were excluded because these phonation types can result in unreliable acoustic measurements. (Davidson, 2019; Patman, Foulkes, & McDougall, 2025) Vowels immediately followed by /r/ were also excluded because rhoticity can substantially alter formant frequencies (Heid & Hawkins, 2000). Out of 5,427 vowel occurrences, 1,265 occurrences were excluded in this process, leaving a total of 4,162 occurrences to analyze.

A Python (Python Software Foundation, 2023) script was used to extract the start and end time of each vowel, as well as the fundamental frequency and first three formants at the midpoint

of each vowel. The pitch ceiling was set to 800 Hz and formant ceiling set to 6500 Hz to improve formant estimation for high-pitched infant-directed vocalizations, where increased fundamental frequency can interfere with accurate formant tracking (Shadle et al., 2016; Boersma & Weenik, 2026). Formant frequencies were extracted at the temporal midpoint of each vowel using Praat's Burg LPC algorithm (Anderson, 1974; Burg, 1972).

Seven acoustic measures were analyzed: fundamental frequency (f0), f0 variability, duration, duration variability, global vowel dispersion, and within-category vowel dispersion. f0 was extracted from Praat at the midpoint of each vowel, and its variability was calculated as the coefficient of variation (CV) of the midpoint f0 values in Hz for each speaker, vocalization target, vocalization type, and vowel category (/i/, /u/, or /ɑ/). Vowel duration was extracted from Praat, and its variability was calculated as the coefficient of variation (CV) of vowel duration in seconds for each speaker, vocalization target, vocalization type, and vowel category. CV was used to normalize variability measures relative to each speaker's mean, allowing variability to be compared across speakers with different baseline pitch and duration characteristics. f0 and vowel duration were positively skewed, so they were natural log-transformed prior to statistical analysis to better satisfy the assumptions of linear mixed-effects modeling. Estimated marginal means were transformed back to the original units (Hz and seconds, respectively) for reporting and visualization. Vowel space area (VSA) was calculated for each speaker, vocalization target, and vocalization type by using the average F1 and F2 values of the three corner vowels (/i/, /u/, and /ɑ/). The area of the triangular vowel space formed by these vowels was calculated using the polygon method described by Kuhl et al (1997).

Global vowel dispersion was calculated as the mean Euclidean distance between each vowel category mean and the centroid of the three-vowel space (/i/, /u/, and /ɑ/) for each speaker,

vocalization target, and vocalization type. Within-category vowel dispersion was calculated as the mean Euclidean distance between each vowel token and the centroid of its corresponding vowel category (e.g., /ɑ/ tokens relative to the /ɑ/ centroid). Both measures were calculated using raw F1 and F2 values and are reported in Hz, with larger values indicating greater dispersion and smaller values indicating greater compactness or consistency within the vowel space.

Prior to statistical analysis, phonetically implausible values were removed. Because automatic formant tracking can produce erroneous measurements, particularly in high-pitched vocalizations, formant values were screened for values outside expected ranges for adult female speakers (Vorperian & Kent, 2007). All vowel tokens with F1 values lower than 150 Hz or greater than 1200 Hz, or with F2 values lower than 500 Hz or greater than 3500 Hz, were excluded. Additionally, /i/ tokens with F2 values lower than 1500 Hz and /u/ tokens with F2 values greater than 2500 Hz were removed because these values were inconsistent with expected acoustic properties of these vowels (Vorperian & Kent, 2007) and were likely attributable to formant tracking errors. This step removed a total of 167 tokens. After this, the Interquartile Range (IQR) method was used to exclude statistical outliers from analysis. First, the data was grouped by vocalization target, vocalization type, and vowel category. Then, all values that fell outside of 1.5 times the interquartile range were excluded. This step removed a total of 85 tokens. The total of 252 tokens removed throughout these processes eliminated 6% of the data, with the final dataset consisting of 3910 vowel tokens. Table 1 illustrates the mean number of vowels produced by the mothers for each vowel category and condition.

Table 1

Mean Count of Vowels for Each Vowel Category, Vocalization Target, & Vocalization Type

	Adult-Directed (AD)		Infant-Directed (ID)	
Vowel	Speech, *M (SD)*	Song, *M (SD)*	Speech, *M (SD)*	Song, *M (SD)*
/i/	30.40 (6.36)	22.47 (8.14)	34.33 (11.8)	21.27 (6.42)
/u/	19.40 (3.16)	20.00 (7.32)	25.93 (7.32)	20.27 (2.09)
/ɑ/	17.53 (4.60)	14.53 (5.40)	20.53 (6.76)	14.00 (3.07)

Prior to plotting and statistical analysis, item-level measurements (i.e., fundamental frequency, duration, and token-level dispersion) were averaged for each participant, vocalization target, vocalization type, and vowel category to facilitate interpretation of the models. Statistical analyses were conducted in R (R Core Team, 2026). Linear mixed-effects models were fit using the lme4 package (Bates et al., 2015), significance tests were performed using lmerTest (Kuznetsova et al., 2017) and estimated marginal means were computed using emmeans (Lenth & Piaskowski, 2026). Graphs were created using the ggplot2 software package (Wickam, 2016). Separate linear mixed-effects models were fit for each dependent variable (mean f0, f0 variability, mean duration, duration variability, vowel space area, global vowel dispersion, and within-category vowel dispersion). Models for f0 and duration were fit using natural log-transformed values, whereas variability and formant-based measures were analyzed in their original units. Vocalization target (adult versus infant), vocalization type (speech versus song), and their interaction were entered as fixed effects, with speaker included as a random intercept. For within-category vowel dispersion, vowel category (/i/, /u/, and /ɑ/) was also entered as a fixed effect to determine whether certain vowels exhibited grater within-category dispersion than

others. Significance of the fixed effects was evaluated using Type III analyses of variance with Satterthwaite's approximation for degrees of freedom. Estimated marginal means (EMMs) were obtained to further interpret significant model effects and visualize adjusted group means.

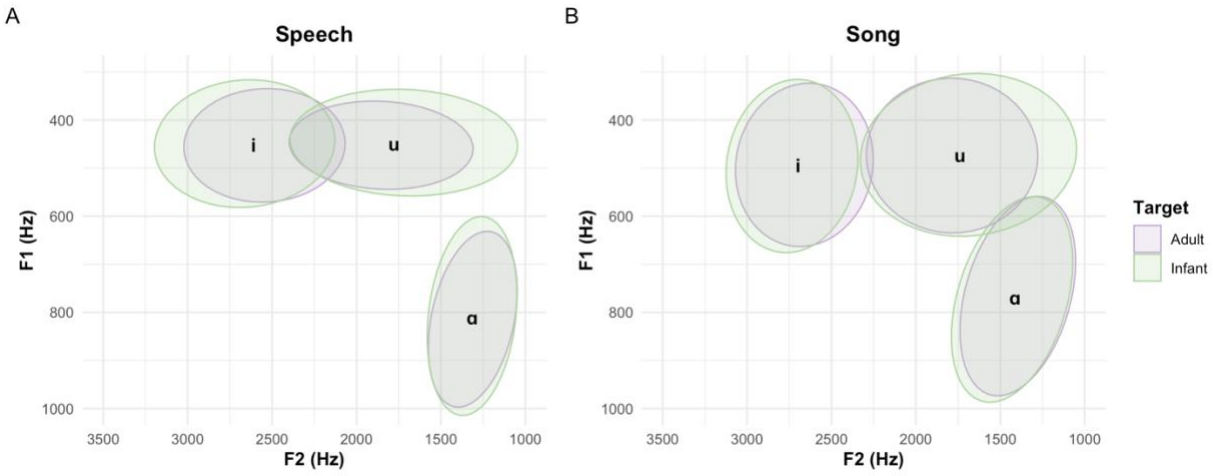
Results

Vowel Space

The phonR (McCloy, 2016) package was used to plot the vowel space of all four conditions averaged by speaker in Figure 3. Like Audibert & Falk (2018), an expansion on F1 and compression on F2 appear to be associated with infant-directedness. Additionally, for each vowel, there appears to be more within-category vowel variability in both speech and song, which supports Martin et al.'s (2015) finding that phonetic contrasts are less clearly produced in ID than AD speech.

Figure 3

Speaker-Averaged Vowel Space of /i/, /u/, and /a/ Across Vocalization Targets & Types



Note. Ellipses represent 1 standard deviation in the bivariate space (68% confidence interval).

Statistical Analysis

Table 2 summarizes the results of the Type III analyses of variance for each model conducted. Notably, effects of vocalization target were larger than those of vocalization type in measures of vowel space (global dispersion, within-category dispersion, and overall vowel space area), while effects of vocalization type were larger than those of vocalization target in measures of fundamental frequency (f0) and duration.

Table 2

Type III Analysis of Variance Results for Linear Mixed-Effects Models of Acoustic Measures

Outcome	Effect	*F*(df)	*p*
Fundamental Frequency (f0)	Target	23.84 (1, 162)	<.001***
	Type	350.63 (1, 162)	<.001***
	Target x Type	5.56 (1, 162)	.020*
f0 Variability	Target	1.90 (1, 162)	.170
	Type	42.82 (1, 162)	<.001***
	Target x Type	7.13 (1, 162	.008**
Duration	Target	10.48 (1, 162)	.001**
	Type	1275.56 (1, 162)	<.001***
	Target x Type	2.02 (1, 162)	.158
Duration Variability	Target	0.83 (1, 162)	.365
	Type	13.15 (1, 162)	<.001***
	Target x Type	2.13 (1, 162)	.146
Global Dispersion	Target	11.61 (1, 42)	.001**
	Type	1.84 (1, 42)	.182
	Target x Type	0.48 (1, 42)	.490
Token Dispersion	Target	9.37 (1, 160)	.003**
	Type	1.41 (1, 160)	.236
	Label	49.58 (2, 160)	<.001***

	Target x Type	0.49 (1, 160)	.483
Vowel Space Area	Target	14.14 (1, 42)	<.001***
	Type	0.09 (1, 42)	.766
	Target x Type	0.02 (1, 42)	.883

*$p \leq .05$, **$p \leq .01$, ***$p \leq .001$.

Distributions of the acoustic measures across vocalization target and type are presented in Figures 4-6. Box plots indicate the median and interquartile range, and points represent individual observations.

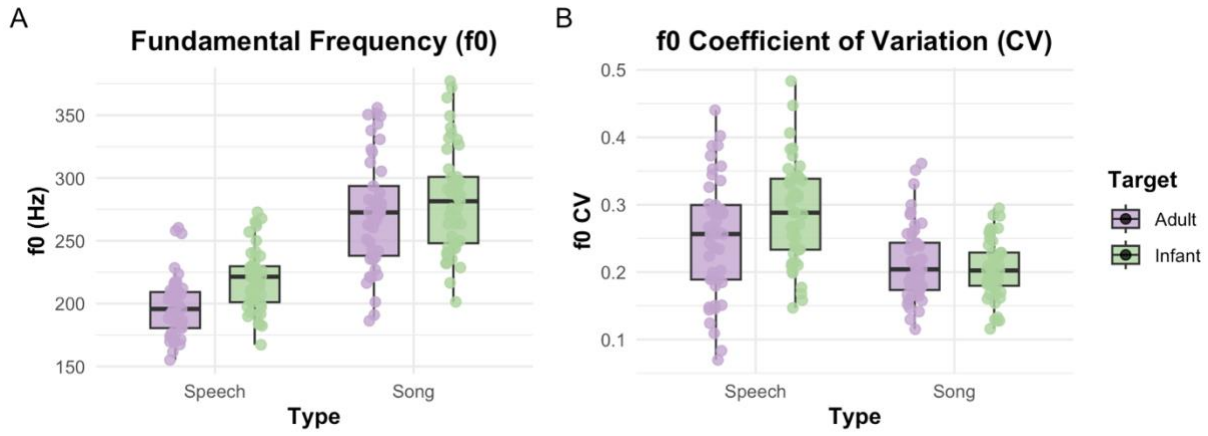
Fundamental Frequency (f0)

Fundamental frequency (f0), which corresponds acoustically to perceived pitch, was examined to determine whether vocalization target and type influenced the pitch of vowel productions. The distribution of f0 and f0 variability is presented in Figure 4.

Figure 4

Distribution of Fundamental Frequency (f0) in Hz and f0 Variability Across Vocalization Targets and Types

.



There was a significant main effect of target $F(1, 162) = 23.84$, $p < .001$ and type $F(1, 162) = 350.63$, $p < .001$. The interaction between target and type was also significant $F(1, 162) = 5.56$, $p = .020$ indicating that the effect of vocalization target on f0 differed significantly between speech and song. Follow-up estimated marginal means (EMMs) comparisons indicated that f0 was significantly higher for ID than AD vocalizations and for songs than speech (all Tukey-adjusted ps $\leq .002$). Averaged across song and speech, mean f0 was estimated to be 247 Hz in ID vocalizations and 230 Hz in AD vocalizations. Averaged across ID and AD vocalizations, mean f0 was estimated to be 273 Hz in song and 207 Hz in speech. The difference between AD and ID vocalizations was significantly smaller in song than in speech. As seen in Figure 4A, ID speech (218 Hz) was approximately 22 Hz higher than AD speech (196 Hz), whereas ID song (279 Hz) was approximately 11 Hz higher than AD song (268 Hz).

f0 Variability.

The model for f0 variability revealed there was no significant main effect of target $F(1, 162) = 1.90$, $p = .170$, but there was a significant main effect of type $F(1, 162) = 42.82$, $p < .001$. Follow-up Tukey-adjusted pairwise comparisons of EMMs indicated that, when averaged across ID and AD vocalizations, f0 variability was 0.25 CV units in speech and 0.19 CV units in song. There was also a significant interaction between target and type $F(1, 162) = 7.13$, $p = .008$, indicating that the effect of vocalization target on f0 variability differed significantly between speech and song. ID speech had significantly greater f0 variability than AD speech ($p = .024$), and speech exhibited greater f0 variability than song for both AD ($p = .034$) and ID ($p < .001$) vocalizations. AD and ID songs did not differ significantly in f0 variability ($p = .798$). As illustrated in Figure 4B, f0 variability was higher in ID speech (0.29 CV) compared to AD speech (0.25 CV), a difference of 0.04 CV units. In contrast, f0 variability was slightly lower in

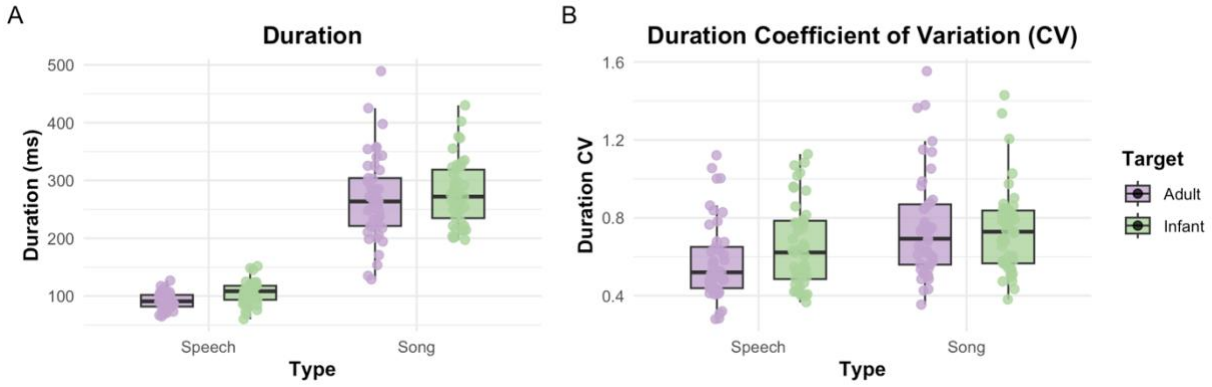
ID song (0.20 CV) than in AD song (0.22 CV), a difference of 0.02 CV units. Therefore, the increase in f0 variability associated with ID vocalizations was present in speech but not in song.

Duration

Vowel duration was examined to determine whether vocalization target and type influenced the temporal characteristics of vowel productions. The distribution of vowel duration and duration variability across vocalization target and type is presented in Figure 5.

Figure 5

Distribution of Vowel Duration in Milliseconds on the Log Scale and Duration Variability Across Vocalization Targets and Types



There was a significant main effect of target $F(1, 162) = 10.50$, $p = .001$ and type $F(1, 162) = 1275.56$, $p < .001$. However, the interaction between target and type was not significant $F(1, 3898.3) = 0.02$, $p = .884$. Follow-up EMMs comparisons indicated that duration was significantly higher for infant-directed than adult-directed vocalizations and for songs than speech across conditions (all Tukey-adjusted ps $\leq .002$). When back-transforming values to seconds, EMMs showed that durations were approximately 168 milliseconds (ms) for ID

vocalizations and 153 ms for ID vocalizations, a difference of 15 ms. A larger difference was seen between speech and song: vowel durations were approximately 97 ms for speech and 266 ms for song, a difference of 169 ms (Figure 5A).

Duration Variability.

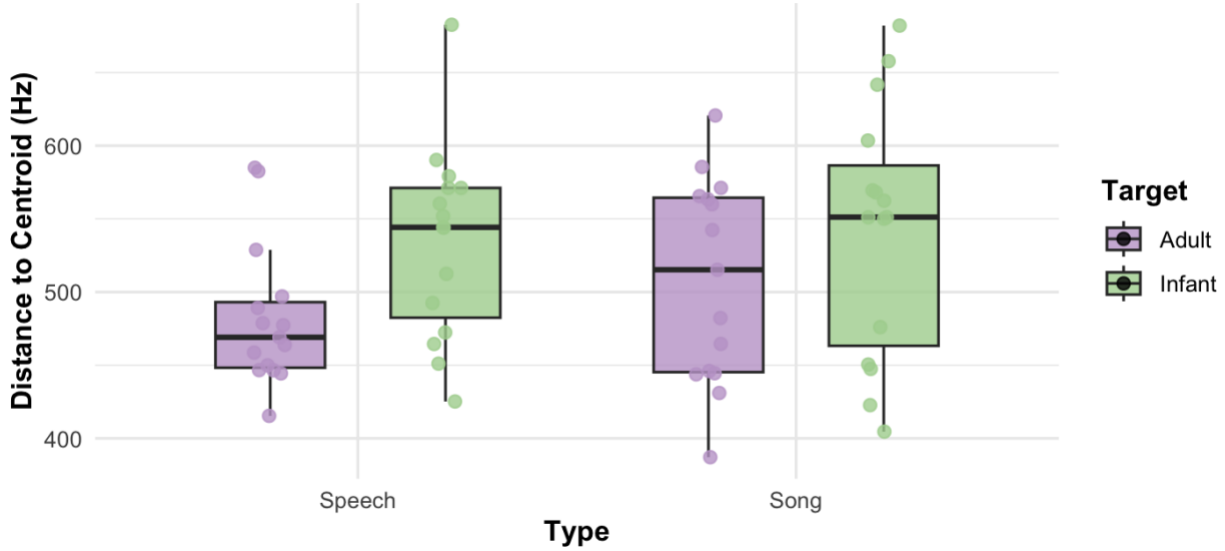
There was no significant main effect of target $F(1, 162) = 0.83$, $p = .365$, but there was a significant main effect of type $F(1, 162) = 13.15$, $p < .001$. The interaction between target and type was not significant $F(1, 162) = 2.13$, $p = .145$. Follow-up EMMs comparison indicated that there was significantly higher duration variability in song than in speech ($p < .001$), with duration variability ranging from 0.62 in speech to 0.74 in song, a difference of 0.12 CV units (Figure 5B).

Global Dispersion

Global vowel dispersion was examined to characterize the distribution of vowel categories within the overall vowel space. This measure was calculated as the mean Euclidean distance of each vowel token from the speaker-specific vowel space centroid, with larger values indicating greater separation of vowel categories within the acoustic space. The distribution of global vowel dispersion across vocalization target and type is presented in Figure 6.

Figure 6

Distribution of Global Vowel Centroid Dispersion in Hz Across Vocalization Targets and Types



There was a significant main effect of target $F(1, 42) = 7.10$, $p = .011$, but neither the main effect of type $F(1, 42) = 1.84$, $p = .182$ nor the interaction between target and type were significant $F(1, 42) = 0.01$, $p = .947$. Follow-up EMMs comparison indicated that there was significantly higher centroid dispersion in infant-directed vocalizations compared to adult-directed vocalizations ($p = .011$), with mean centroid dispersion of approximately 538 Hz for infant-directed vocalizations and approximately 495 Hz for adult-directed vocalizations, a difference of 43 Hz. There was also significantly lower centroid dispersion in song compared to speech ($p = .014$), with mean centroid dispersion of approximately 508 Hz for song and 525 Hz for speech (Figure 6).

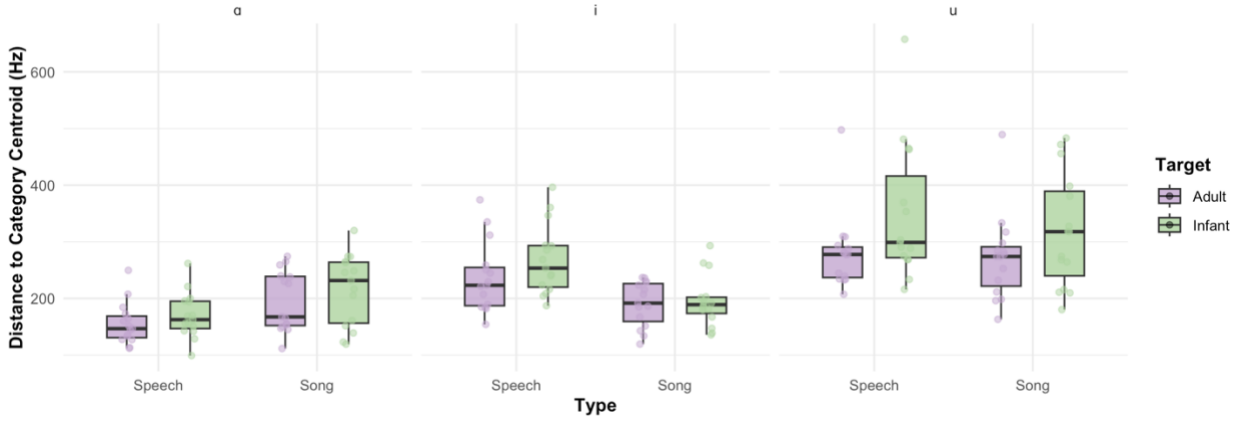
Within-Category Dispersion

Within-category vowel dispersion was examined to quantify variability among individual productions within each vowel category. This measure was calculated as the mean Euclidean

distance between each vowel token and the centroid of its corresponding vowel category (e.g., all /i/ productions relative to the /i/ centroid). Larger values indicate greater within-category variability in vowel productions. The distribution of within-category vowel dispersion across vocalization target, vocalization type, and vowel category is presented in Figure 7.

Figure 7

Distribution of Within-Category Vowel Dispersion in Hz Across Vowel Categories, Vocalization Targets, and Vocalization Types



There was a significant main effect of target $F(1, 160) = 9.37$, $p = .002$ and vowel $F(2, 160) = 49.58$, $p < .001$. However, the main effect of type $F(1, 160) = 1.41$, $p = .236$ was not significant, nor was the interaction between target and type $F(1, 3895.7) = 2.65$, $p = .104$. Follow-up EMMs comparison indicated that there was significantly higher within-category dispersion in infant-directed vocalizations compared to adult-directed vocalizations ($p = .002$), with mean within-category dispersion of 253 Hz for infant-directed vocalizations and 221 Hz for adult-directed vocalizations, a difference of 32 Hz. Additionally, there were significant differences between token dispersion for each vowel (all ps <.01), with /u/ observed to have the

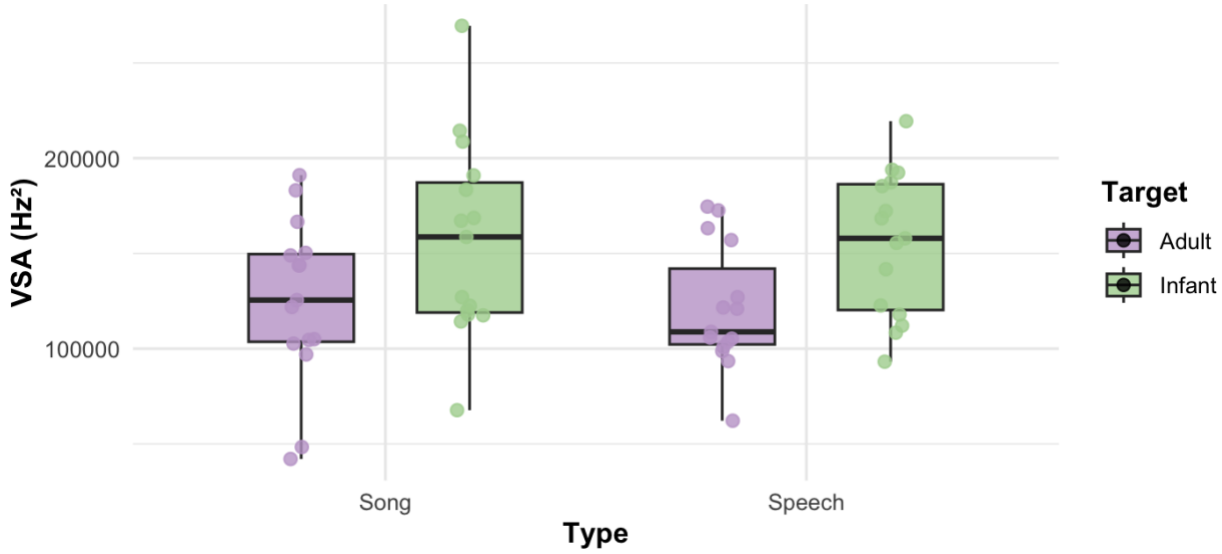
greatest mean within-category dispersion of approximately 305 Hz, /ɑ/ observed to have the lowest mean within-category dispersion of approximately 184 Hz, and /i/ observed to have a mean within-category dispersion of approximately 223 Hz (Figure 7).

Vowel Space Area

Vowel space area (VSA) was measured to examine how vocalization target and type affect the overall vowel space. Figure 8 presents the distribution of VSA across vocalization targets and types.

Figure 8

Distribution of VSA in Hz^2 Across Vocalization Targets, and Vocalization Types

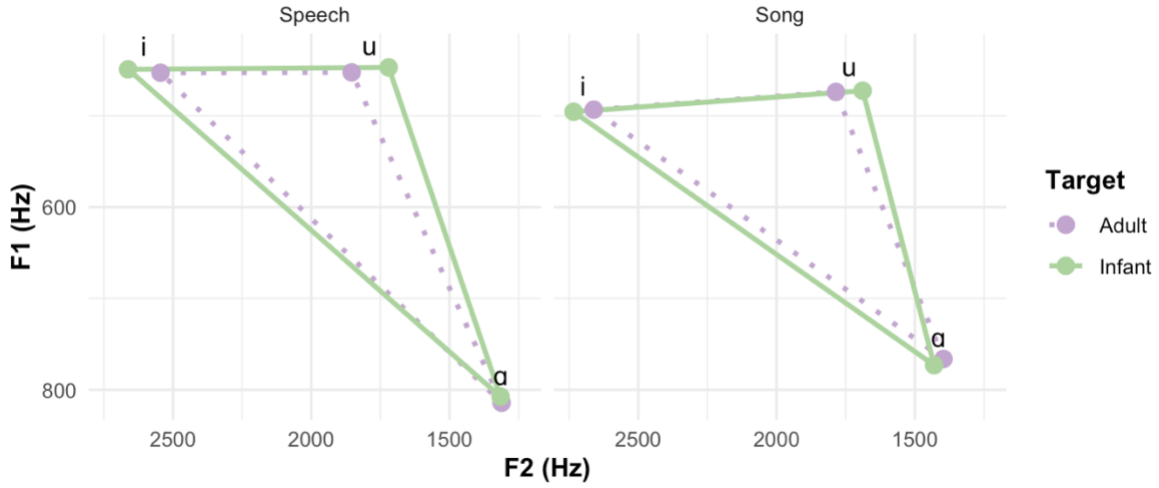


There was a significant main effect of target $F(1, 42) = 14.13$, $p < .001$. However, there was no significant main effect of type $F(1, 42) = 0.09$, $p = .766$ and the interaction between target and type was not significant, $F(1, 42) = 0.02$, $p = .883$. Follow-up estimated marginal means comparison indicated that there was significantly higher VSA in infant-directed

vocalizations compared to adult-directed vocalizations ($p < .001$), with infant-directed vocalizations exhibiting a mean VSA of approximately 155,908 Hz^2 and adult-directed vocalizations exhibiting a mean VSA of approximately 123,018 Hz^2. Figure 8 visualizes the varying VSAs across the four conditions.

Figure 8

Vowel Space Area in Hz^2 by Vocalization Target & Type



Exploratory Analyses

Speech.

Since speech was elicited in two different contexts (scripted versus naturalistic), exploratory analyses were conducted to determine if and how these different elicitation contexts influenced the acoustic characteristics of the vowels analyzed. Scripted speech differed from naturalistic speech for several measures, including f0, f0 variability, vowel duration, and duration variability. However, elicitation context did not significantly interact with vocalization target for these measures. Measures of vowel space did not differ significantly between elicitation contexts, with ID speech exhibited greater vowel space area and global dispersion than AD

speech in both scripted and naturalistic contexts. A significant target-by-elicitation interaction was observed for within-category vowel dispersion, indicating that the difference between ID and AD speech depended on elicitation context. Follow-up Tukey-adjusted EMMs comparisons indicated that dispersion in ID naturalistic speech (256 Hz) was 74 Hz higher than dispersion in AD naturalistic speech (182 Hz), while dispersion in ID scripted speech (238 Hz) was only 16 Hz higher dispersion than AD scripted speech (222 Hz). These findings suggest that differences in within-category vowel dispersion ID and AD speech were more pronounced in naturalistic than in scripted speech. However, the only two pairwise comparisons that reached significance were the contrasts between ID scripted speech and AD naturalistic speech ($p = .017$) and between ID naturalistic speech and AD naturalistic speech ($p < .001$).

Song.

Since six different songs sorted into two unique groups (ABC: *Mamma Mia, Take Me Out to the Ball Game,* and *Here Comes the Sun;* DEF: *Hallelujah, Dancing Queen,* and *You've Got a Friend in Me*) were elicited, exploratory analyses were conducted to examine whether acoustic measures differed between the two groups of songs. Significant main effects of song group were observed for f0, f0 variability, vowel duration, and vowel space area, whereas no significant effects of song group were found for global or within-category vowel dispersion. Song group did not significantly interact with vocalization target for most measures, suggesting that the differences between ID and AD singing were generally consistent across songs. A significant target-by-song interaction was observed for f0 variability, indicating that the effect of vocalization target differed between songs. Follow-up Tukey-adjusted comparisons indicated that DEF songs exhibited greater f0 variability than ABC songs, particularly in adult-directed vocalizations. There was a difference of 0.05 CV units between songs ABC (0.05 CV) and DEF

(0.10 CV) in AD vocalizations, but a difference of only 0.02 CV units between songs ABC (0.06 CV) and DEF (0.08 CV). This suggests that differences in f0 variability between song groups were attenuated during ID singing compared to AD singing. The only pairwise comparisons that reached significance were the contrasts between AD songs ABC and DEF ($p < .001$), and between ID songs ABC and AD songs DEF ($p = .013$).

Discussion

Question 1: How do (1) vowel fundamental frequency (f0), (2) vowel f0 variability, (3) vowel duration, (4) vowel duration variability, (5) global vowel dispersion, (6) within-category vowel dispersion, and (7) vowel space area differ in AD versus ID vocalizations across both speech and song?

It was hypothesized that ID vocalizations would have (1) higher f0, (2) greater f0 variability, (3) longer vowel duration, (4) greater vowel duration variability, (5) greater global vowel dispersion, (6) greater within-category vowel dispersion, and (7) greater vowel space area (VSA) compared to AD vocalizations. As expected, vowel f0 was found to be significantly higher in ID vocalizations than AD vocalizations, consistent with the well-established trend in the literature. Estimated marginal means (EMMs) indicated that ID speech was approximately 20 Hz higher than AD speech, and ID song was approximately 13 Hz higher than AD song. However, no overall significant differences were observed in f0 variability between the two vocalization targets. This contradicts the hypothesis of the present study and the findings of several other studies (Cox et al., 2022; Neer et al., 2025; Benders, St. George, & Fletcher, 2021).

A similar pattern was observed in duration. Vowel duration was found to be significantly higher in ID compared to AD vocalizations. EMMs indicated that vowels were, on average, approximately 14 ms longer in infant-directed vocalizations (177 ms) than in adult-directed

vocalizations (153 ms). The finding regarding vowel duration is in line with previous research finding longer vowel durations in ID compared to AD vocalizations (Cox et al., 2022; Rosslund et al., 2024), and it corroborates the results of Audibert & Falk (2018). However, its variability did not significantly differ between the two vocalization targets.

Global vowel dispersion was found to be significantly higher in ID compared to AD vocalizations. EMMs indicated that global dispersion was, on average, 43 Hz greater in ID vocalizations (538 Hz) compared to AD vocalizations (495 Hz), indicating that vowel categories were produced farther from the speaker's vowel-space centroid. This is contrary to the findings of Audibert & Falk (2018) and Cox et al. (2023) but corroborates the findings of other research (Cox et al., 2022; Hartman, Ratner, & Newman, 2016; Kuhl et al., 1997). It a possibility that the language spoken has an impact on global vowel dispersion, as Audibert & Falk analyzed German vocalizations and Cox et al. (2023) analyzed Danish vocalizations. However, methodological differences between studies may have also played a role, as infants were physically present in Audibert & Falk's paper, and Cox et al. conducted a meta-analysis rather than an experimental study. Within-category vowel dispersion was also found to be significantly higher in ID than AD vocalizations, replicating Audibert & Falk's results, with estimated marginal means indicating that within-vowel dispersion was, on average, 32 Hz greater in ID vocalizations (253 Hz) than in AD vocalizations (221 Hz), indicating that individual vowel tokens were produced farther from the centroid for their vowel category. Finally, as hypothesized, VSA was observed to be significantly higher in ID compared to AD vocalizations. EMMs indicated that VSA was, on average, 32,899 Hz^2 greater in ID vocalizations (155,907 Hz^2) than in AD vocalizations (123,018 Hz^2), reflecting a larger vowel space, which is consistent with the pattern observed throughout the literature (Cox et al., 2022; Hartman, Ratner, & Newman, 2016; Kuhl et al., 1997).

Overall, the results indicate that ID vowels are characterized by greater f0, duration, global dispersion, within-category dispersion, and vowel space area compared to AD vowels. However, variability in f0 and duration did not significantly differ in ID vowels compared to AD vowels.

Question 2: How do these same variables differ between speech and song across both AD and ID vocalizations?

It was hypothesized that song would have (1) higher vowel f0, (2) lower vowel f0 variability, (3) longer vowel duration, (4) lower vowel duration variability, (5) lower global vowel dispersion, (6) lower within-category vowel dispersion, and (7) lower vowel space area compared to speech. As predicted, f0 was significantly higher in song than in speech, with EMMs indicating that song increased f0 by approximately 61 Hz compared to speech in ID vocalizations, and by 72 Hz compared to speech in AD vocalizations. Likewise, f0 variability was significantly lower in song compared to speech, with EMMs indicating 0.07 CV units lower variability in song (CV units = 0.20) than speech (CV units = 0.27), consistent with previous findings of more stable f0 in AD song (Ozaki et al., 2024) and lower pitch variability in ID song compared to ID speech by Tsang, Falk, & Hessel (2016). The higher f0 and lower f0 variability observed in song are likely attributable to its fixed melodic structure, which constrains speakers' ability to modulate pitch compared to speech.

Both duration and its variability found to be significantly higher in song compared to speech, supporting only the hypothesis about duration itself. Duration was found to be, on average, 169 ms longer in song (266 ms) than in speech (97 ms). This result is in line with previous work finding longer syllable duration (Raposo de Madeiros & Cabral, 2018) and vowel duration in AD song (Hansen, Bokshi, & Khorram, 2020) and ID song (Falk et al., 2021; Falk &

Audibert, 2021). Duration variability was, on average, 0.12 CV units higher in song (0.74 CV) than in speech (0.62 CV). This counters previous findings of lower rhythmic variability in ID song (Falk et al., 2021; Alviar et al., 2022), however, this may be due to methodological differences. Alviar et al. (2022) measured duration variability using the Normalized Pairwise Variability Index (nPVI), which examines contrast between adjacent events in a sequence and Falk et al. (2021) used standard deviation, which is sensitive to outliers and the scale of the data (e.g., a larger mean will generally co-occur with a larger standard deviation). This study used the coefficient of variation (CV) to measure duration variability because it provides a global measure of dispersion, unlike nPVI, which examines neighboring durations. Additionally, the use of CV facilitates comparison of durations relative to each speaker's mean duration, minimizing the influence of differences in baseline duration that affect standard deviation.

Surprisingly, no significant differences in global dispersion, within-category dispersion, or overall vowel space area were found between speech and song. This contradicts previous work finding a more compressed vowel space in AD song compared to speech (Bradley, 2018; Audibert & Falk, 2018) and lower within-category vowel dispersion in song (Audibert & Falk, 2018). The small sample size of this study may have reduced the power needed to detect a significant effect.

Overall, these results demonstrate that sung vowels were characterized by higher f0, lower f0 variability, and greater duration and duration variability compared to spoken vowels. In contrast, no significant differences in vowel space measures were observed between sung and spoken vowels. Furthermore, the effects of vocalization type were larger for measures of f0 and duration than the effects of vocalization target, whereas vowel space measures showed stronger effects of vocalization target than vocalization type. These findings suggest that vocalization

target and vocalization type influence distinct aspects of vowel production, with vocalization target primarily affecting vowel articulation and vocalization type primarily affecting pitch and timing. Exploratory analyses further supported this interpretation, as differences between elicitation contexts and individual songs primarily affected pitch and timing measures rather than vowel-space measures.

Question 3: Is there an interaction effect between vocalization target (AD versus ID) and vocalization type (speech versus song) on each of the dependent variables?

Significant interaction effects between vocalization target and vocalization type were found only for f0 and f0 variability. The effect of infant-directedness on f0 was smaller in song than in speech: infant-directed speech was approximately 22 Hz higher than adult-directed speech (218 vs. 196 Hz), whereas infant-directed song was only about 11 Hz higher than adult-directed song (279 vs. 268 Hz). A different pattern was observed for f0 variability. Infant-directedness was associated with significantly greater f0 variability in speech, but no significant difference was observed between AD and ID song. In fact, f0 variability was numerically lower in ID compared to AD song. One possible explanation for these reduced ID-AD differences in song is that melodic structure limits opportunities for additional pitch modulation. Songs require speakers to follow predetermined pitch trajectories, which may reduce flexibility in modifying both mean pitch and pitch variability based on vocalization target. The decreased f0 variability seen in ID song compared to AD is surprising given that caregivers are known to increase their f0 variability when addressing infants (Cox et al., 2022).

Limitations & Future Directions

There are several important limitations of this study to note. First, the fact that infants and adults were not physically present during the time of recording may have influenced the acoustic

modulations made by the mothers who participated. However, the observation of significant differences between AD and ID conditions suggests that mothers were able to produce vocalizations with acoustic characteristics associated with ID communication despite the absence of their infant. Additionally, post-hoc analyses indicated that some acoustic variables differed between the naturalistic and scripted speech conditions, as well as between the different songs. These differences may have introduced additional variability into the dataset and influenced the observed effects. Future research may further examine the differences between ID speech produced in different context (i.e., naturalistic/conversational speech versus scripted/read speech). Furthermore, differences in melodic and rhythmic structure across songs may have contributed to variability in acoustic measures, making it difficult to disentangle effects of vocalization type from properties of individual songs. Future research should examine differences in acoustic measures between individual songs to elucidate the impact of song structure more clearly. Another limitation is that both stressed and unstressed vowels were included in the dataset. Because stress can influence acoustic properties such as duration, fundamental frequency, and vowel quality, effects of vocalization target and type may differ depending on stress pattern. Future work should examine whether stress moderates the effects of vocalization target and vocalization type. Also, the relatively small sample size of 15 participants may have limited statistical power, particularly for detecting smaller effects and interactions. Future analyses incorporating the remaining participants may provide greater insight into acoustic differences between vocalization conditions. Finally, the sample consisted of monolingual English speakers in the greater Seattle area, limiting the generalizability of the results to other populations and linguistic contexts. Future research examining these patterns

across different languages, populations, and developmental stages with larger speaker samples will further clarify the acoustic characteristics of AD and ID speech and song.

Conclusion

The present study examined acoustic differences between vowels produced in adult-directed (AD) and infant-directed (ID) speech and song. Specifically, this study investigated fundamental frequency (f0), f0 variability, duration, duration variability, global dispersion, within-category dispersion, and vowel space area (VSA) to better understand how mothers modify their vocalizations when speaking and singing to their infants compared to an adult loved one.

Taken together, the findings of this study demonstrate that infant-directedness and vocalization type affect some shared and some different acoustic characteristics of vowels. ID vocalizations were characterized by increased mean f0, longer vowel durations, and expanded vowel spaces. This is consistent with Falk & Audibert's (2021) finding that infant-directedness significantly affected more acoustic characteristics than other dimensions (e.g., speech versus song). However, the effects of vocalization target on measures of f0 and duration were lower than the effects of vocalization type on those measurements, suggesting that infant-directedness is more strongly associated with changes in the vowel space than with f0 and duration. Song was associated with increased mean f0 and longer vowel duration, and uniquely characterized by reduced f0 variability and increased duration variability, reflecting the influence of musical structure on vocalizations. Importantly, interactions between vocalization target and type were found only for f0 and f0 variability. The increase in f0 and f0 variability associated with infant-directedness was much smaller in song than in speech, suggesting that acoustic modulations

characteristic of infant-directedness may be constrained when vocalizations are embedded within song.

The present findings contribute to a burgeoning understanding of the acoustic characteristics that distinguish ID song from both AD song and ID/AD speech and highlight the importance of considering speech and song as distinct contexts in which mothers communicate with their infants. Although both ID speech and ID song serve important social, communicative, and educational functions for infants, the present results suggest that mothers modify acoustic characteristics differently depending on the different constraints imposed by speech versus song.

References

Alviar, C., Sahoo, M., Edwards, L. A., Jones, W., Klin, A., & Lense, M. (2023). Infant-directed song potentiates infants' selective attention to adults' mouths over the first year of life. *Developmental Science*, *26*(5), e13359. https://doi.org/10.1111/desc.13359

Andersen, N. (1974). On the calculation of filter coefficients for maximum entropy spectral analysis. *Geophysics*, *39*(1), 69-72. https://doi.org/10.1190/1.1440413

Audibert, N., & Falk, S. (2018, June). Vowel space and f0 characteristics of infant-directed singing and speech. In Proceedings of the 19th international conference on speech prosody (Vol. 153, p. 157). ISCA.

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). *WhisperX: Time-Accurate Speech Transcription of Long-Form Audio* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2303.00747

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Boersma, Paul & Weenink, David (2025). Praat: doing phonetics by computer [Computer program]. Version 6.4.41, retrieved 28 August 2026 from https://praat.org

Bradley, E. D. (2018). A comparison of the acoustic vowel spaces of speech and song. *Linguistic Research*, 35(2), 381–394. https://doi.org/10.17250/khisli.35.2.201806.006

Burg, J. P. (1972). The relationship between maximum entropy spectra and maximum likelihood spectra. Geophysics, 37(2), 375-376. https://doi.org/10.1190/1.1440265

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

Corbeil, M., Trehub, S. E., & Peretz, I. (2013). Speech vs. singing: Infants choose happier sounds. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00372

Cox, C., Bergmann, C., Fowler, E., Keren-Portnoy, T., Roepstorff, A., Bryant, G., & Fusaroli, R. (2022). 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

Daniele, J., (2017). A Tool for the Quantitative Anthropology of Music: Use of the nPVI Equation to Analyze Rhythmic Variability within Long-term Historical Patterns in Music. *Empirical Musicology Review* 11(2), 228-233. https://doi.org/10.18061/emr.v11i2.4893

Davidson, L. (2019). The effects of pitch, gender, and prosodic context on the identification of creaky voice. *Phonetica*, *76*(4), 235-262. https://doi.org/10.1159/000490948

Delavenne, A., Gratier, M., & Devouche, E. (2013). Expressive timing in infant-directed singing between 3 and 6 months. *Infant Behavior and Development*, *36*(1), 1–13. https://doi.org/10.1016/j.infbeh.2012.10.004

Falk, S., & Audibert, N. (2021). Acoustic signatures of communicative dimensions in codified mother-infant interactions. *The Journal of the Acoustical Society of America*, *150*(6), 4429–4437. https://doi.org/10.1121/10.0008977

Falk, S., Fasolo, M., Genovese, G., Romero-Lauro, L., & Franco, F. (2021). Sing for me, Mama! Infants' discrimination of novel vowels in song. *Infancy*, *26*(2), 248–270. https://doi.org/10.1111/infa.12387

Falk, S., & Kello, C. T. (2017). Hierarchical organization in the temporal structure of infant-direct speech and song. *Cognition*, *163*, 80–86. https://doi.org/10.1016/j.cognition.2017.02.017

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

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., De Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, *16*(3), 477–501. https://doi.org/10.1017/S0305000900010679

Grieser, D. L., & Kuhl, P. K. (1988). Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese. *Developmental Psychology*, *24*(1), 14–20. https://doi.org/10.1037/0012-1649.24.1.14

Gros-Louis, J., West, M. J., Goldstein, M. H., & King, A. P. (2006). Mothers provide differential feedback to infants' prelinguistic sounds. *International Journal of Behavioral Development*, *30*(6), 509–516. https://doi.org/10.1177/0165025406071914

Hartman, K. M., Ratner, N. B., & Newman, R. S. (2016). Infant-directed speech (IDS) vowel clarity and child language outcomes. *Journal of Child Language*, *44*(5), 1140–1162. https://doi.org/10.1017/s0305000916000520

Heid, S., & Hawkins, S. (2000, May). An acoustical study of long-domain/r/and/l/coarticulation. In *Proceedings of the 5th seminar on speech production: Models and data* (pp. 77-80). Kloster Seeon Germany.

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*, *6*(11), 1545–1556. https://doi.org/10.1038/s41562-022-01410-x

Kendall, Tyler and Erik R. Thomas. 2010. Vowels: vowel manipulation, normalization, and plotting in R. R package. cran.r-project.org/web/packages/vowels/index.html

Kent, R. D., & Vorperian, H. K. (2018). Static measurements of vowel formant frequencies and bandwidths: A review. *Journal of Communication Disorders*, *74*, 74–97. https://doi.org/10.1016/j.jcomdis.2018.05.004

Kitamura, C., Thanavishuth, C., Burnham, D., & Luksaneeyanawin, S. (2001). Universality and specificity in infant-directed speech: Pitch modifications as a function of infant age and sex in a tonal and non-tonal language. *Infant Behavior and Development*, *24*(4), 372–392. https://doi.org/10.1016/S0163-6383(02)00086-3

Kjeldsen, C. P., Neel, M. L., Jeanvoine, A., & Maitre, N. L. (2025). Investigation of mothers' elicited infant-directed speech and singing for preterm infants. *Pediatric Research*, *97*(7), 2367–2375. https://doi.org/10.1038/s41390-024-03618-1

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

Kuznetsova A., Brockhoff PB., & Christensen RHB. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13), 1-26. https://doi.org/10.18637/jss.v082.i13

Lenth R, Piaskowski J (2026). _emmeans: Estimated Marginal Means, aka Least-Squares Means. https://doi.org/10.32614/CRAN.package.emmeans, R package version 2.0.2, https://CRAN.R-project.org/package=emmeans

Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. Language, 48(4), 839-862.

Longhi, E. (2009). `Songese': Maternal structuring of musical interaction with infants. *Psychology of Music*, *37*(2), 195–213. https://doi.org/10.1177/0305735608097042

Martin, A., Schatz, T., Versteegh, M., Miyazawa, K., Mazuka, R., Dupoux, E., & Cristia, A. (2015). Mothers speak less clearly to infants than to adults: A comprehensive test of the hyperarticulation hypothesis. *Psychological Science*, *26*(3), 341–347. https://doi.org/10.1177/0956797614562453

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. (2017). Montreal Forced Aligner: trainable text-speech alignment using Kaldi. In Proceedings of the 18th Conference of the International Speech Communication Association.

McCloy, D. R. (2016). phonR: tools for phoneticians and phonologists. R package version 1.0-7.

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

Neer, E. M., Brahmbhatt, A., Walsh, C. R., & Warlaumont, A. S. (2025). Pitch characteristics of real-World infant-directed speech vary with pragmatic context, perceived adult gender, and infant gender. *PLOS One*, *20*(6), e0326569. https://doi.org/10.1371/journal.pone.0326569

Ozaki, Y., Tierney, A., Pfordresher, P. Q., McBride, J. M., Emmanouil Benetos, Polina Proutskova, Chiba, G., Liu, F., Jacoby, N., Purdy, S. C., Opondo, P., W. Tecumseh Fitch, Hegde, S., Rocamora, M., Thorne, R., Nweke, F., Sadaphal, D. P., Sadaphal, P. M., Shafagh Hadavi, … Savage, P. E. (2024). Globally, songs and instrumental melodies are slower and higher and use more stable pitches than speech: A registered report. *Science Advances*, *10*(20). https://doi.org/10.1126/sciadv.adm9797

Patman, C., Foulkes, P., & McDougall, K. (2025). Acoustic methods for analysing breathy and whispery voices: A systematic review. *Phonetica*, *82*(4), 271–299. https://doi.org/10.1515/phon-2025-0007

Raposo de Medeiros, B., & Cabral, J. P. (2018, June). Acoustic distinctions between speech and singing: Is singing acoustically more stable than speech? In Proceedings of the 19th international conference on speech prosody (Vol. 153, p. 542). ISCA.

Reddy, S., and Stanford, J. 2015. A Web Application for Automated Dialect Analysis. In Proceedings of NAACL-HLT 2015.

Rosenfelder, I., Fruehwald, J., Keelan Evanini, Seyfarth, S., Gorman, K., Prichard, H., & Jiahong Yuan. (2015). FAVE: Speaker fix (Version v1.2.2) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.22281

Rosslund, A., Mayor, J., Mundry, R., Singh, A. P., Cristia, A., & Kartushina, N. (2024). A longitudinal investigation of the acoustic properties of infant-directed speech from 6 to 18 months. *Royal Society Open Science*, *11*(11). https://doi.org/10.1098/rsos.240572

Shadle, C. H., Nam, H., & Whalen, D. H. (2016). Comparing measurement errors for formants in synthetic and natural vowels. *The Journal of the Acoustical Society of America*, *139*(2), 713-727. https://doi.org/10.1121/1.4940665

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-Directed Speech Facilitates Word Segmentation. *Infancy*, *7*(1), 53–71. https://doi.org/10.1207/s15327078in0701_5

Trainor, L. J., Clark, E. D., Huntley, A., & Adams, B. A. (1997a). 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

Trainor, L. J., Clark, E. D., Huntley, A., & Adams, B. A. (1997b). 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://doi.org/10.1111/cdev.12647

Van Der Klis, A., Adriaans, F., & Kager, R. (2023). Infants' behaviours elicit different verbal, nonverbal, and multimodal responses from caregivers during early play. *Infant Behavior and Development*, *71*, 101828. https://doi.org/10.1016/j.infbeh.2023.101828

Werker, J. F., Pegg, J. E., & McLeod, P. J. (1994). A cross-language investigation of infant preference for infant-directed communication. *Infant Behavior and Development*, *17*(3), 323–333. https://doi.org/10.1016/0163-6383(94)90012-4

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4. https://ggplot2.tidyverse.org.