Speech Prosody 2026
26-29 May 2026, Philadelphia, Pennsylvania, USA

On the Relationship between Breathing and Prosody in the CAPIL Corpus

Melissa A. Redford, Bess Frerichs, Kristopher Kyle

Department of Linguistics, University of Oregon, Eugene, OR 97403-1290 USA
redford@uoregon.edu, bfrerich@uoregon.edu, kkyle2@uoregon.edu

Abstract

The relationship between breathing and prosody is well established, though its nature remains unclear: do the physiological constraints of breathing help define prosodic patterning, or do communicative imperatives behind prosodic patterning drive breathing behavior? Assuming a functional relationship exists, the current investigation adopts a developmental approach to clarify its direction using the soon-to-be-available Child and Adult Phrasing, Inhalation, and Language (CAPIL) corpus. The corpus consists of about 18 hours of spontaneous speech from 30 5-year-old children and 30 20-year-old adults, segmented into pause-speech and inhalation-exhalation intervals. Effects of prior inhalation depth on subsequent phrasing, pitch change, and pitch reset are tested against effects of age, speaking style, and utterance length or boundary strength. Comparison between linguistic-only models and breathing-inclusive models (controlling for prior exhalation depth) indicated significant relationships between inhalation depth and the prosodic variables, but not pitch reset. The effect of inhalation depth on phrasing interacted with age group, with a similar trend for pitch change: the breathing-prosody relationship was stronger in adults than in children. The developmental findings suggest maturation of motor speech planning, consistent with the communicative-imperative account of the breathing-prosody relationship.

Index Terms: speech production, motor speech development, prosodic acquisition

1. Introduction

The rhythm of breathing has long been argued to provide a biological foundation for prosodic phrasing [1], [2]. Similarly, the relationship between lung volume and subglottal pressure has been proposed as a biological basis for prosodically-coded pitch (f0) changes across utterances and for utterance-initial pitch height [3], [4]. In this study, we use the Child and Adult Phrasing, Inhalation, and Language (CAPIL) corpus to gain insight into the breathing-prosody relationship. The corpus, currently being prepared for public access through TalkBank [5], includes approximately 18 hours of spontaneous speech in two speaking styles (narrative and expository) from kindergarten-aged children and young adults. Simultaneous recording of speech acoustics and breathing kinematics during elicitation enables us to examine the coordination of breathing and language during speaking from a developmental perspective.

Breathing during speaking differs substantially from breathing at rest. In healthy adults, the expiratory phase of the respiratory cycle during speech is 9 times as long as the inspiratory phase, whereas during rest the phases maintain a more nearly 1-to-1 relationship [6]. Inhalation depth also varies substantially with the length of the upcoming message [6], even when starting lung volume is statistically or experimentally controlled using previous exhalation depth [7], [8]. These clear adaptations to communicative imperatives are cortically managed, reflecting speech motor control over breathing. However, speech breathing is also constrained by the essential life-sustaining gas-exchange function of breathing and thus by physiological factors such as lung capacity. Since both speech motor control and lung capacity are immature in young children, the speech breathing-prosody relationship should differ between child and adult speech. The direction of these differences can provide insight into whether the breathing-prosody relationship is driven primarily by communicative imperatives or by physiological constraints: a weaker relationship in children than in adults would suggest immature motor control and thus a communicative-imperative account, whereas a stronger relationship would suggest a physiological-constraints account.

The analyses reported here test for a breathing-prosody relationship in the CAPIL corpus and for age-related differences in this relationship. Our analyses are based on three utterance-level prosodic variables, where an utterance is defined as a breath group, with each breath group consisting of speech produced during a single exhalation cycle preceded by an inhalation. The prosodic variables under study are (1) prosodic phrasing; (2) pitch change; and (3) pitch reset. Linguistic-only versus breathing-inclusive models of these variables were built to explain the CAPIL data. The breathing-prosody relationship predicts a significant difference between the two models, with the breathing-inclusive model explaining more variance in the data than the linguistic-only model. Insights into the direction of the breathing-prosody relationship are gained by examining the strength and slope of inhalation and exhalation depth effects on the prosodic variables by age group.

2. Methods

2.1. Participants

Thirty children (16 female) and 30 adults (15 female) were recruited to participate in a study on the coordination of language and breathing during speaking. Children ranged in age from 5;1 (years;months) to 6;10 (M = 5;9, SD = 5 months); adults ranged in age from 18;0 to 24;5 (M = 20;7, SD = 18 months). All were native speakers of a standard variety of West Coast American English. The majority of children and adults self-identified as racially/ethnically white-only (N = 23 and N = 18, respectively). Three children and four adults reported a history of speech delay/disorders, though none were in speech therapy. All children and adults had expressive vocabularies that were typical for their age, as determined using the Expressive Vocabulary Test [9]. Participants were compensated for their time at the rate of $15/hour.

2.2. Speech Elicitation

All 60 speakers narrated two short Claymation videos in real-time (∼5 min each); hey watched each video in full before narrating it on a second viewing. All speakers also provided expository speech samples in response to descriptive and explanatory prompts (e.g., ”Tell me what you do on a typical day”; ”How

10.21437/SpeechProsody.2026-20

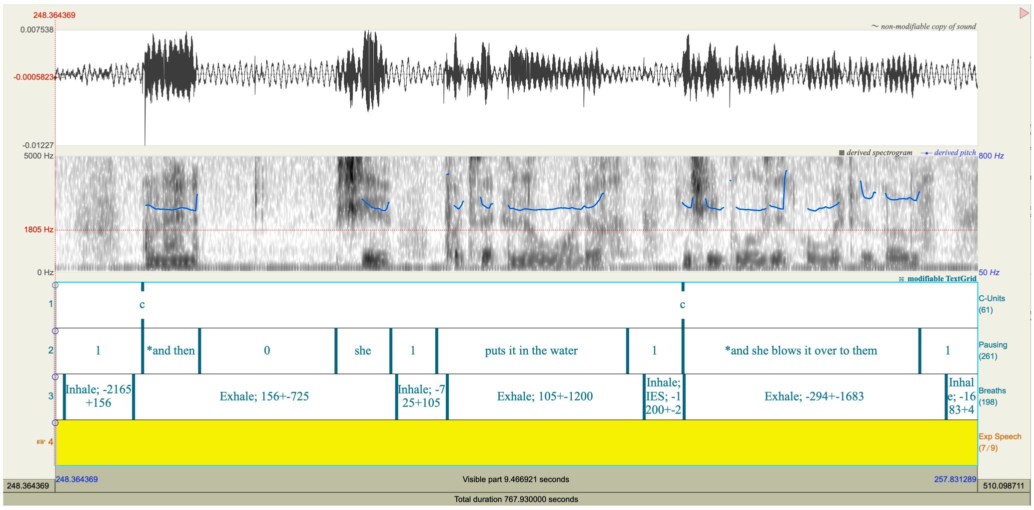

Figure 1: *Speech acoustic and breathing waveforms were independently segmented into pause-delimited utterances and inhalation-exhalation intervals, respectively. Speech was transcribed and annotated for C-unit boundaries (Loban, 1976). Silent pauses were coded for the presence or absence of inhalation (1/0). Kinematically-identified inhalation and exhalation intervals were also incorporated into the TextGrids along with starting and ending amplitude values (in volts).*

do you play your favorite game?"), speaking for 3 to 5 minutes per prompt with an experimenter interlocutor asking open-ended follow-up questions.

For breathing kinematic segmentation, respiratory inductance plethysmography (RIP; Inductotrace, Ambulatory Monitoring Inc.) captured chest and abdominal wall movements via fabric transducer bands placed around the abdomen at the navel and chest under the armpits; the kinematic signal was sampled at the same frequency as a simultaneously recorded directional microphone audio signal (44,100 sps) and then downsampled to 100 sps for analysis. The gain from the chest and abdominal channels was adjusted so that their movements contributed equally to the signal. For acoustic segmentation and analyses, speakers wore a lavalier microphone and audio was separately digitally recorded using the Zoom F1 Field Recorder.

2.3. Segmentation and Coding

All speech bouts in the narrative and expository speech samples were segmented into inhalation and exhalation intervals using the kinematic signal in conjunction with the time-locked speech waveform. Inhalations were marked from a local minimum to a local maximum, and exhalations from the local maximum to a local minimum. The separately recorded acoustic data were pause-delimited and transcribed in Praat [10] according to CHAT [11] and lab-specific conventions. Communication units (C-units) [12]—stretches of speech that convey a complete thought—were marked in the transcribed text following Systematic Analysis of Language Transcripts (SALT) conventions [13]. Pauses in the Praat-transcribed data were marked for the presence or absence of inhalation based on the independent kinematic segmentation. The independently determined onset/offset times for inhalation/exhalation intervals and amplitude values (in volts) at these timepoints were added as a tier in the TextGrids. Experimenter speech was identified on a final tier. Figure 1 illustrates the acoustic segmentation and TextGrid annotation, including the incorporated breathing tier.

2.4. Data Extraction

Once segmented, transcribed and coded, speech samples were automatically segmented by breath group (BG) to create utterance-length audio files. The procedure resulted in 14,825 files for acoustic/prosodic analysis (adults: $N = 6{,}669$; children: $N = 8{,}156$). Fifty-nine percent of the utterance files were from the narrative speech samples (adults: 60.7% from narratives; children: 57.7% from narratives); the remainder were from expository speech samples. The individual utterance files were then acoustically/prosodically analyzed using the Momel and INTSINT algorithms [14]. These complementary algorithms provide a language-independent analysis of intonation by modeling the continuous pitch contour as a series of tonal targets. We used these algorithms to extract prosodically relevant pitch information, including the absolute f0 values of the first and last INTSINT-identified tones for each utterance as well as the minimum, maximum, and mean f0 across all tones. This information, along with linguistic boundary and breathing kinematic information extracted from the TextGrids themselves, provided the data for analyses. These data were cleaned as follows.

All utterances that were marked as containing experimenter speech were excluded from the analyses (N = 2,536) to ensure that the prosodic information extracted was in fact the speakers' speech. The majority of these exclusions were children's utterances (N = 2,333) and associated with expository speech (N = 1762), when the experimenter was often required to ask open-ended follow-up questions to elicit a sufficiently long speech sample. We also excluded from analysis utterances with preceding silent pauses that included multiple inhalations (N=743) because previous BG exhalation depth was our proxy for starting lung volume in analyses with pre-utterance inhalation depth as a predictor. Again, the majority of these exclusions were children's utterances (N = 567). They were mainly associated with narrative speech (N = 687) when speakers would sometimes pause to simply watch action unfold in the video. Finally, we excluded all utterances for which the amplitude of the associated inhalation or exhalation was marked in the kinematic segmentations as unreliable based on a set of criteria established using the AV-recordings in conjunction with the kinematic recordings (N = 2091). Once again, the majority of these exclusions were children's utterances (N = 1498) and associated with narrative speech (N = 1221). After exclusions, a total of 9,455 utterances (i.e., breath groups) remained for analysis (adults: N = 5,697; children: N = 3,758).

2.5. Analyses

Analyses compared linguistic-only to breathing-inclusive models of three prosodic variables associated with utterances: prosodic phrasing; pitch change; and pitch reset. Prosodic phrasing was operationalized as the number of phrases per utterance, calculated as the number of within-utterance silent pauses$+1$. Given its discrete count distribution, phrasing was modelled using a Poisson generalized linear mixed-effects model (GLMM). Pitch change was operationalized as the f0 range in Hertz of INTSINT-identified tones within the utterance and was modelled using linear mixed effects. Pitch reset was the difference in Hertz of the absolute f0 value associated with the first INTSINT-identified tone minus that associated with the last INTSINT-identified tone from the previous utterance and was likewise modelled using linear mixed effects.

Age group, speaking style, and either number of words per utterance or pre-utterance boundary status (C-unit or not) were the fixed effects in the linguistic-only models: models of phrasing and pitch change used number of words per utterance given length effects on these variables; models of pitch reset used pre-utterance boundary status given that pitch reset is a boundary phenomenon. Pre-utterance inhalation depth and the exhalation depth of the preceding utterance were additional fixed effect predictors in the breathing-inclusive models. These breathing predictors were z-scored within speaker. Random intercept (speaker) and slope (style) models were constructed. If a maximal random effects model failed to converge, it was simplified to a random intercept-only model. Models were fit on utterances with complete data for all predictors (phrasing and pitch reset: N = 7,941; pitch change: N = 6,587, after additionally excluding utterances with maximum INTSINT-identified f0 exceeding 500 Hz as likely octave-doubling errors). All models were built in R using lme4 package [15], with the MuMIn package [16] for model comparison. Model comparisons are reported with associated χ^2 and p-values. Marginal and conditional R^2 values (R^2_m and R^2_c) for the best-fit models are also reported. We used the lmerTest package [17] to investigate fixed effects of interest in the best-fit models. β, SE, t- or z-values, and p-values are presented for these effects. Figures were created using the ggplot2 package [18].

3. Results

The modeling results are presented by prosodic variable below. Comparison between linguistic-only and breathing-inclusive models indicated significant relationships between breathing and both prosodic phrasing and pitch change, with a very small but significant effect of previous exhalation depth on pitch reset. The effect of breathing on prosodic phrasing interacted significantly with age group: the breathing-prosody relationship was stronger in adults than in children.

3.1. Phrasing

Models of prosodic phrasing were maximal random effects models. The linguistic-only model included age group, speaking style, and number of words per utterance and their interactions as fixed effects. The breathing-inclusive model added pre-utterance inhalation depth and preceding exhalation depth as additional fixed effects. Model comparison revealed a significant effect of breathing on phrasing, $\chi^2(2) = 23.36$, $p < .001$. Adding the 2-way age group by breathing interaction further improved the model fit over the breathing-inclusive main-effects model, $\chi^2(2) = 13.63$, $p < .001$. Even by the conservative calculations of the Poisson GLMM, the fixed effects in this final best-fit model explained 11% of the variance in phrasing ($R^2_m = .108$ and $R^2_c = .144$), with the effect of pre-utterance inhalation depth stronger in adult speech than in child speech, $\beta = -0.065$, $SE = 0.020$, $z = -3.31$, $p < .001$, as shown in Figure 2. Phrasing also varied systematically with utterance length, $\beta = 0.347$, $SE = 0.053$, $z = 6.54$, $p < .001$ (see Figure 3).

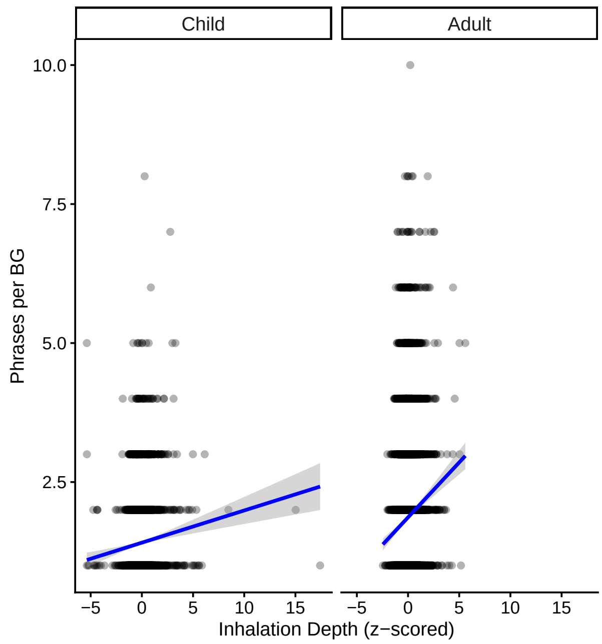

Figure 2: *Phrasing (N per BG) varies systematically with pre-utterance inhalation depth, albeit differently across age groups.*

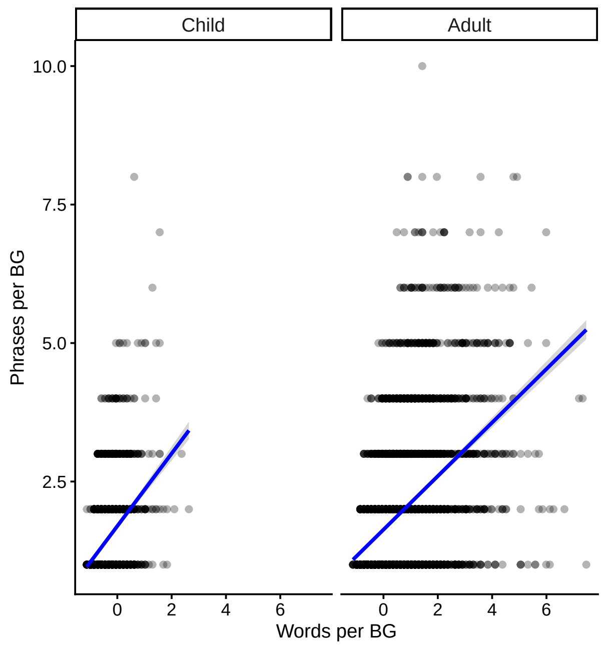

Figure 3: *Phrasing (N per BG) varies systematically with utterance length (Words per BG) and by age group.*

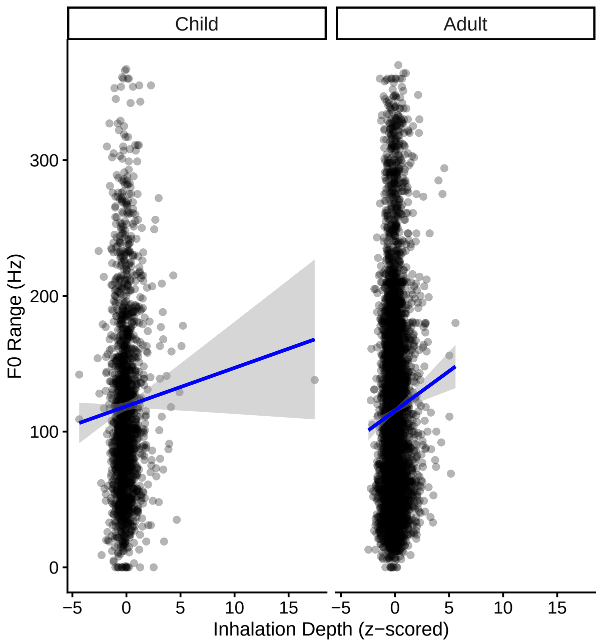

Figure 4: *Pitch change (f0 range) varies systematically with pre-utterance inhalation depth.*

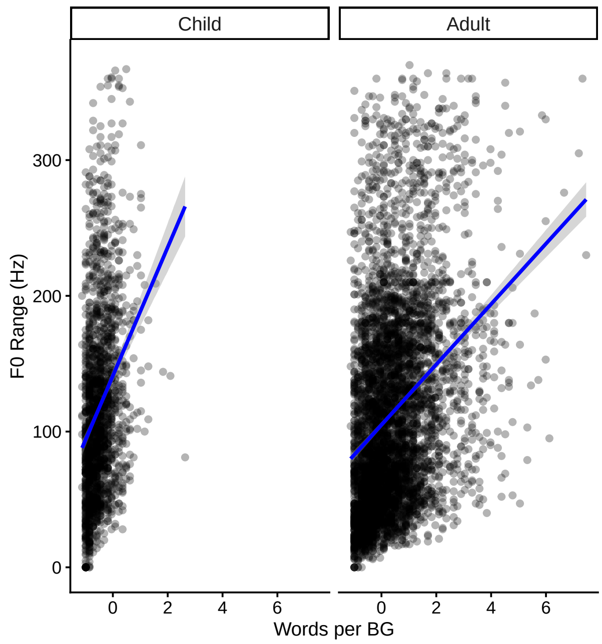

Figure 5: *Pitch range (i.e., change) varies systematically with utterance length (Words per BG) and by age group.*

3.2. Pitch Change

Models of pitch change were also maximal random effects models. The linguistic-only model again included age group, speaking style, and number of words per utterance and their interactions as fixed effects. The breathing-inclusive model added pre-utterance inhalation depth and preceding exhalation depth as additional fixed effects. Model comparison revealed a significant effect of breathing on pitch change, $\chi^2(2) = 6.37$, $p < .05$. The two-way breathing by age group interactions did not improve model fit. The best-fit breathing-inclusive model explained 26% of the variance in pitch change, though the fixed effects explained just 9% of this variance ($R^2_m = .089$ and $R^2_c = .260$). As shown in Figure 4, deeper pre-utterance inhalations were associated with larger pitch changes than shallower pre-utterance inhalations, $\beta = 2.59$, $SE = 1.02$, $t = 2.52$, $p < .05$. The extent of pitch change also varied systematically with utterance length, $\beta = 49.56$, $SE = 6.31$, $t = 7.87$, $p < .001$ (see Figure 5).

3.3. Pitch Reset

Models of pitch reset were random-intercept only models with age group, speaking style, and linguistic boundary included as fixed effects in the linguistic-only model. Pre-utterance inhalation depth and preceding exhalation depth were added as additional fixed effects in the breathing-inclusive model. Model comparison indicated a significant but very small effect of breathing on pitch reset, $\chi^2(2) = 7.19$, $p < .05$, $\Delta R^2_m < .001$. Within the breathing-inclusive model, preceding exhalation depth was a significant predictor ($\beta = 3.58$, $SE = 1.34$, $t = 2.68$, $p < .01$), while pre-utterance inhalation depth was not ($\beta = -1.04$, $SE = 1.53$, $t = -0.68$, $p = .496$). Pitch reset also varied systematically with age group ($\beta = 16.51$, $SE = 5.04$, $t = 3.28$, $p < .01$), speaking style ($\beta = -7.94$, $SE = 2.88$, $t = -2.76$, $p < .01$), and linguistic boundary ($\beta = -9.39$, $SE = 2.76$, $t = -3.40$, $p < .01$). The negative C-unit boundary effect on pitch reset was unexpected: utterances ended with rising intonation, particularly at C-unit boundaries, indicating that CAPIL speech samples are characterized by uptalk. This inference was supported when pre- and post-boundary tone f0 were individually examined.

4. Conclusion

This study introduced the CAPIL corpus to investigate the breathing-prosody relationship in spontaneous speech and whether developmental differences can clarify the nature of this relationship. Analysis of nearly 8,000 breath groups from kindergarten-aged children and young adults revealed that inhalation depth significantly predicted both prosodic phrasing and pitch change, though not pitch reset. Critically, the effect of inhalation was weaker in children than in adults when the interaction between age group and the inhalation/exhalation depth predictors improved model fit. This developmental finding provides insight into the nature of the breathing-prosody relationship: if physiological constraints drove this relationship, speakers with smaller lung capacities (children) should show tighter coupling between breathing capacity and prosodic patterning, producing a stronger relationship, not a weaker one. Instead, the weaker effect of inhalation on phrasing in children points to immature motor control, with the breathing-prosody relationship strengthening as speakers develop the ability to strategically coordinate the respiratory and linguistic/prosodic systems to meet communicative goals. These results demonstrate the value of CAPIL for taking a developmental approach to address theoretical questions about speech production mechanisms that underpin the coordination of breathing and language during speaking.

5. Acknowledgements

This work was supported by grant #BCS-2315232 from the National Science Foundation (PI: Redford).

6. References

[1] A. Henderson, F. Goldman-Eisler, and A. Skarbek, "Temporal patterns of cognitive activity and breath control in speech," *Language and Speech*, vol. 8, no. 4, pp. 236–242, 1965.

[2] P. Lieberman, *Intonation, Perception, and Language*. Cambridge, MA: MIT Press, 1966.

[3] J. Vaissière, "Perception of intonation," in *The Handbook of Speech Perception*, D. B. Pisoni and R. E. Remez, Eds. Oxford, UK: Blackwell, 2005, pp. 236–263.

[4] P. J. Watson, A. H. Ciccia, and G. Weismer, "The relation of lung volume initiation to selected acoustic properties of speech," *J. Acoust. Soc. Am.*, vol. 113, no. 5, pp. 2812–2819, 2003.

[5] B. MacWhinney, "Understanding spoken language through TalkBank," *Behav. Res. Methods*, vol. 51, no. 4, pp. 1919–1927, 2019.

[6] S. Fuchs and A. Rochet-Capellan, "The respiratory foundations of spoken language," *Annu. Rev. Linguist.*, vol. 7, no. 1, pp. 13–30, 2021.

[7] S. Fuchs, C. Petrone, J. Krivokapić, and P. Hoole, "Acoustic and respiratory evidence for utterance planning in German," *J. Phon.*, vol. 41, no. 1, pp. 29–47, 2013.

[8] D. H. Whalen and J. M. Kinsella-Shaw, "Exploring the relationship of inspiration duration to utterance duration," *Phonetica*, vol. 54, no. 3–4, pp. 138–152, 1997.

[9] K. T. Williams, *Expressive Vocabulary Test*, 3rd ed. Bloomington, MN: NCS Pearson, 2019.

[10] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]," Version 6.4.13, 2025. [Online]. Available: `https://praat.org`

[11] B. MacWhinney, "Tools for analyzing talk part 1: The CHAT transcription format," *Carnegie*, vol. 16, 2017.

[12] W. Loban, *Language Development: Kindergarten through Grade Twelve*. NCTE Committee on Research Report No. 18, 1976.

[13] J. Heilmann and J. F. Miller, "Systematic analysis of language transcripts solutions: A tutorial," *Perspect. ASHA Spec. Interest Groups*, vol. 8, no. 1, pp. 1–18, 2023.

[14] D. J. Hirst, "A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation," in *Proc. 16th Int. Congr. Phon. Sci. (ICPhS XVI)*, 2007.

[15] D. Bates, M. Maechler, B. Bolker, S. Walker, R. H. B. Christensen, H. Singmann, and M. B. Bolker, "Package 'lme4'," *Convergence*, vol. 12, no. 1, p. 2, 2015.

[16] K. Barton and M. K. Barton, "Package 'mumin'," *Version*, vol. 1, no. 18, p. 439, 2015.

[17] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "Package 'lmertest'," *R Package Version*, vol. 2, no. 0, p. 734, 2015.

[18] H. Wickham, W. Chang, and M. H. Wickham, "Package 'ggplot2'. Create elegant data visualisations using the grammar of graphics," *Version*, vol. 2, no. 1, pp. 1–189, 2016.