Optimizing Random Sampling of Daylong Audio

Anonymous CogSci submission

Abstract

While naturalistic daylong audio recordings of children's auditory environments have the potential to reveal key insights about the input children receive and inform our theories of language development, it also presents various methodological hurdles. In the present work, we used three fully transcribed daylong audio recordings to investigate the challenge of manually extrapolating aggregate statistics and quantify the kinds of sampling choices daylong researchers can make. Our findings highlight sampling choices that maximize sampling from the full distribution of the day and potential tradeoffs between human effort and obtaining accuracy.

Keywords: daylong audio; language development; manual extrapolation; sampling

Introduction

Decades of research has shown that environmental factors play a role in shaping a child's language development (Lieven, 2016), and variation in the language environment can to some degree explain individual variation in development (Goodman, Dale & Li, 2008; Hoff, 2006; Huttenlocher et al., 2010).

Newer advances in technology allow researchers to unobtrusively record longform audio of a child's naturalistic auditory environment. Daylong audio provides a rich dataset that opens the door to a vast range of inquiries; researchers can investigate measures of the overall amount of speech or specifically child-directed speech (Casillas et al., 2017; Weisleder & Fernald, 2013; Weiss et al., 2022), the temporal distribution of select words (Montag, 2020), of infant and adult vocalizations (Warlaumont et al., 2022), of music (Mendoza & Fausey, 2021), and much more. Given the potential for uncovering important facts about children's early language experiences, there is a growing interest in collecting naturalistic corpora of children's auditory environments. These investigations that aim to systematically investigate the patterns and variability of the language environment, including cross-culturally (Casillas et al., 2020, Cychosz et al., 2021), longitudinally (Warlaumont, 2016), and in clinical contexts (McDaniel et al., 2020), not only vitally enrich our understanding of developmental trajectories, but also ultimately have the potential to strengthen our theories of language development.

For these recordings to be useful, researchers must somehow extract features of interest, which presents a major methodological challenge, and often employ either automated or manual approaches. Automated approaches to annotation allow researchers to analyze the entirety of the audio, but only a limited number of features can be assessed and quality is variable. Software such as the widely used Language Environment Analysis (LENA™, the LENA Research Foundation, Boulder, CO) system can exhibit performance quality that is often context dependent (Xu et al., 2009) and has been shown to systematically overestimate certain metrics (Ramirez et al., 2021). Independent assessments highlight that many of its validation studies are lacking in thorough independent peer reviews and call for improved reporting (Cristia et al., 2020). Open-source alternatives, such as ALICE (Räsänen et al., 2021) provide a promising and more accessible option, however these tools are in their early iterations and likewise require thorough evaluation and further development.

Manual transcription and annotation of a daylong audio is a hefty, often years-long process that requires substantial funding and training resources, and thus presents a daunting task to researchers. Instead, researchers can opt for a less laborious path of either transcribing or annotating smaller segments of daylong audio recordings and extrapolate or narrowing the scope of the transcription or annotation process whenever possible (Clemens & Kegel, 2021; Ferjan Ramírez et al., 2022; Fields-Olivieri & Cole, 2022), thus limiting the information that can be analyzed from the recordings.

A wide range of sampling methods have been used: Weisleder and Fernald (2013) manually transcribed 60-minute samples; Ramírez-Esparza et al. (2014, 2017) sampled 30-second intervals; Casillas et al. (2020, 2021) used 1-minute, 2.5-minute and 5-minute samples. Prior reviews have demonstrated that differing sampling methods can misrepresent the daylong distribution of a feature and urge for intentional sampling choices (Bergelson et al., 2019; Tamis-LeMonda et al., 2017). As the use of daylong audio research continues to grow, so too does the need for well-established methods, including sampling for manual extrapolation

In the present work, we aim to address the methodological challenges with manual extrapolation and ask: *How can you most accurately estimate a daylong statistic by randomly sampling from the day?* Working from 3 fully transcribed daylong audio recordings, we investigate how to optimize sampling choices to extrapolate more accurate estimates while minimizing the amount of human annotation effort that is required. We suggest a pipeline that daylong researchers may implement in their sampling methods.

Methods

Sampling simulations were performed upon 3 fully transcribed daylong recordings, and are referred to as Transcripts A, B, and C in this paper. All recordings were collected using LENA recorders within English-speaking homes. Transcript A was collected and transcribed by VanDam (2018) and is publicly available on the HomeBank database (VanDam et al., 2016). Transcripts B and C were collected by Fausey and Mendoza (2018) and transcribed by Montag (2020). All recordings were manually transcribed following ACLEW conventions (Soderstrom et al., 2021).

Figure 1 presents the cumulative speech distribution of each recording. Small gaps in speech indicate moments where the child is napping or temporarily out of the home (Transcript A in Figure 1) and thus no speech was transcribed, whereas large gaps, such as in Transcript C where speech is transcribed only in the morning and evening, indicate that the child was out of the home for most of the day. Other features of the individual transcripts are presented in Table 1.

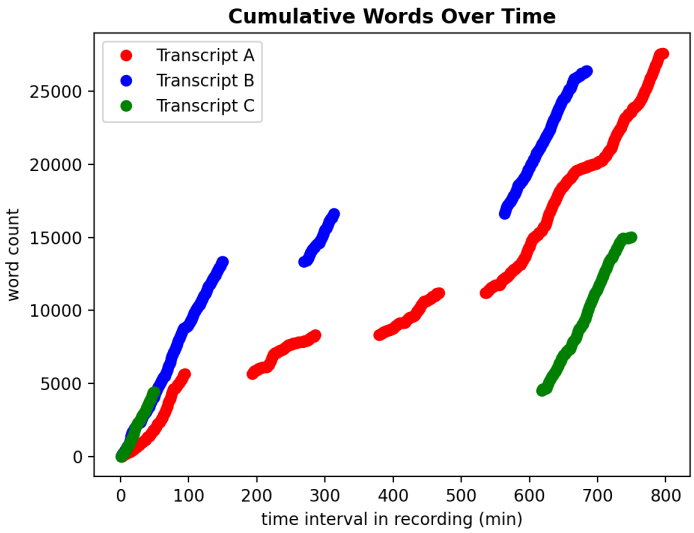

Figure 1: Cumulative speech distribution (per minute) for each transcript.

Table 1: Select transcript features

	Child Age	Recording Length (min)	Total Spoken Time (min)	Total Words
A	1 yr., 7 days	839.95	536.51	27,604
B	10 mo., 9 days	683.32	313.64	26,423
C	11 mo., 7 days	748.32	179.47	15,010

General Sampling Procedure

We manipulated sampling choices along two primary dimensions: (1) the total amount of time sampled per daylong recording and (2) sampling interval size. Sampling more from the day should yield better extrapolated estimates, but at the cost of greater human effort. Likewise, shorter sampling interval sizes rather than larger continuous intervals should improve extrapolation, and the present analyses will probe to what extent, quantitatively, extrapolation can be improved. Another dimension by which sampling methods vary is how and from where sampling intervals are selected. There are many ways in which researchers might sample, such as sampling from particularly dense periods of recording, randomly sampling, or some combination of pre-processing the audio followed by random sampling. We first identified only long intervals of silence, then generated random intervals across the remaining audio. We aimed to quantify the trade-offs between human effort and extrapolation accuracy and understand how different sampling choices contribute to overall accuracy.

A sampling algorithm was implemented using Python, with the intention that this may be a general workflow when working with a raw, untranscribed daylong recording:

1. Identify large intervals of silence (>30 minutes) that indicate the target child is napping/out of the home and should not be sampled from. We found that manually identifying long periods of silence prior to sampling improved our extrapolated estimates.
2. Select a sampling interval size (length of each sample 30 seconds – 60 minutes) and desired total sampled time (i.e., number of total samples, n). For example, if sampling interval size is 10 minutes and desired total sampled time is 100 minutes, then n=10 sampling windows will be randomly selected
3. Generate n random sampling windows that do not overlap into identified intervals of silence
4. Sample words/desired linguistic feature within generated sampling windows
5. Use sampled count to calculate the daylong estimate:

$$\frac{\text{total audio length-total silence intervals}}{\text{total sampled audio time}} \times \text{sampled count}$$

The present work implements this sampling procedure to estimate total word counts, however this method can be implemented to estimate any feature of interest from the environment, such as conversational turns, child-directed speech, select phrases/words, etc.

Variations on the Sampling Method

A challenge that emerges with any sampling method is how to sample utterances that overlap into sampling boundaries (see Figure 2). Different methods of counting, or not counting, utterances that partially fall in sampling boundaries may have consequences for the extrapolated counts. We compared four different sampling methods.

Method 1: "Conservative" Sampling Method 1 only samples from utterances that are fully within the generated interval window and disregards utterances that overlap across the interval (Figure 2, in blue).

Method 2: Include Utterance Past Boundary Method 2 samples from utterances that begin within but end outside the desired sampling interval. To account for words sampled outside the sampling interval, the boundary is extended to the end of the overlapping utterance (Figure 2, in pink). Thus, when extrapolating the daylong estimate, this observed total sampled time is used.

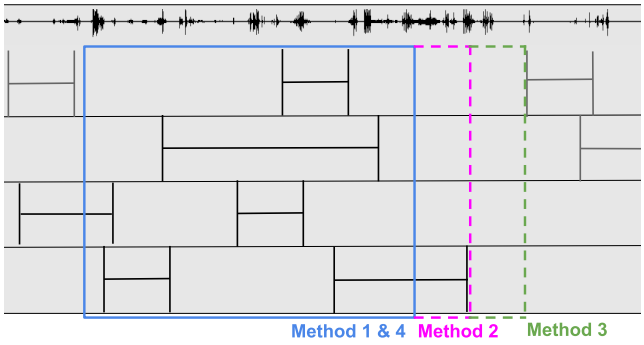

Figure 2: Visualization of audio segmented into utterances and diarized by speaker (ex. "CHI", "MOT", etc.). Different sampling methods treat sampling boundaries that fall within utterances differently. Sampling methods 1 & 4 retain the generated sampling boundary (in blue), while methods 2 and 3 extend the boundary (in pink and green, respectively).

Method 3: Include Utterance Past Interval Boundary & Following Silence In Method 3, utterances that begin within but end outside the desired sampling boundary are sampled from as they are in Method 2. To balance out the proportion of speech and silence within the samples (potential issues extending each random sample with time that only ever includes speech), the sampling interval boundary is extended even further to include any silence following the overlapping utterance before the next utterance (Figure 2, in green). This observed total sampled time is used to extrapolate.

Method 4: Sample ~Half of Overlapping Utterances Method 4 samples from half of the words in utterances that overlap either at the beginning or the end of the desired sampling interval. This is done to approximate which words from the overlapping utterances that may fall into the desired sampling interval boundary. Since this method only considers about half of the words from overlapping utterances, sampling boundaries are not extended to the start or end timestamps of the overlapping utterances (Figure 2, in blue).

Results

Sampling estimates of daylong word counts were obtained using 30-second, 1-minute, 5-minute, 10-minute, 30-minute, and 60-minute sampling interval sizes and a total sampled time ranging from 30 minutes to 120 minutes. Results will present the distribution of estimate accuracy obtained for 100 simulations.

Performance of Sampling Methods

Figure 3 presents the distribution of estimate accuracy across the 4 sampling methods for sampling interval sizes of 30 seconds, 5 minutes, and 60 minutes and a total sampled time of 120 minutes. Many of the observed results are consistent with what one might intuit from logical consequences of sampling but show a magnitude of these logical effects.

The performance across the 4 sampling methods first highlights that the choice of sampling method appears to matter the most at shorter sampling interval sizes (column 1 of Figure 3). Intuitively, this makes sense. These sampling methods aim to address different ways on how to count utterances that overlap across sampling boundaries. A smaller sampling interval, and thus a greater number of samples, will result in more instances of overlapping utterances and a greater opportunity for different sampling methods to affect the results.

Additionally, across the 3 transcripts, the 4th sampling method ("Sample ~Half") seems to produce the most narrow and accurate distributions. Contrastingly, within the shortest sampling interval size, the 3rd sampling method ("Sample Overlapping & Silence") produces the widest range of daylong estimates across the transcripts (the long tails). This also intuitively makes sense; if the length of time between one utterance to the next varies widely and inconsistently in naturalistic speech, then the size of the observed sampling boundary varies widely and inconsistently, and thus does the calculated daylong estimate.

Further, as sampling interval size increases, it is not the choice of sampling method that seems to matter most, but rather the choice of sampling interval size. With longer sampling intervals, there is a greater overlap in the distribution curves of the individual sampling methods (columns 2 and 3 of Figure 3, respectively). However, more notably, there is a greater shift along the x axis further away from a proportion of 1, particularly for Transcripts B and C. Longer sampling intervals seem to over-estimate the total word counts of the day, the reasons for which we will probe in following sections. Overall, the 4th sampling method in which about half of the words from overlapping utterances are sampled seems to perform the best and is the sampling method used to compare choices of sampling interval and total sampled time.

Performance across Sampling Interval Size and Total Sampled Time

It might be inferred without any sampling that shorter sampling interval sizes and greater total time sampled would yield both more accurate and more precise estimates of total counts. These analyses allow us to quantify these effects to make clear recommendations about the trade-off between human transcription effort and extrapolation quality.

Figure 4 presents the estimate accuracy distributions across total sampled time (30 minutes, 60 minutes, 100 minutes, and 120 minutes) and sampling interval sizes (30 seconds to 60 minutes). As expected, across all transcripts and regardless of

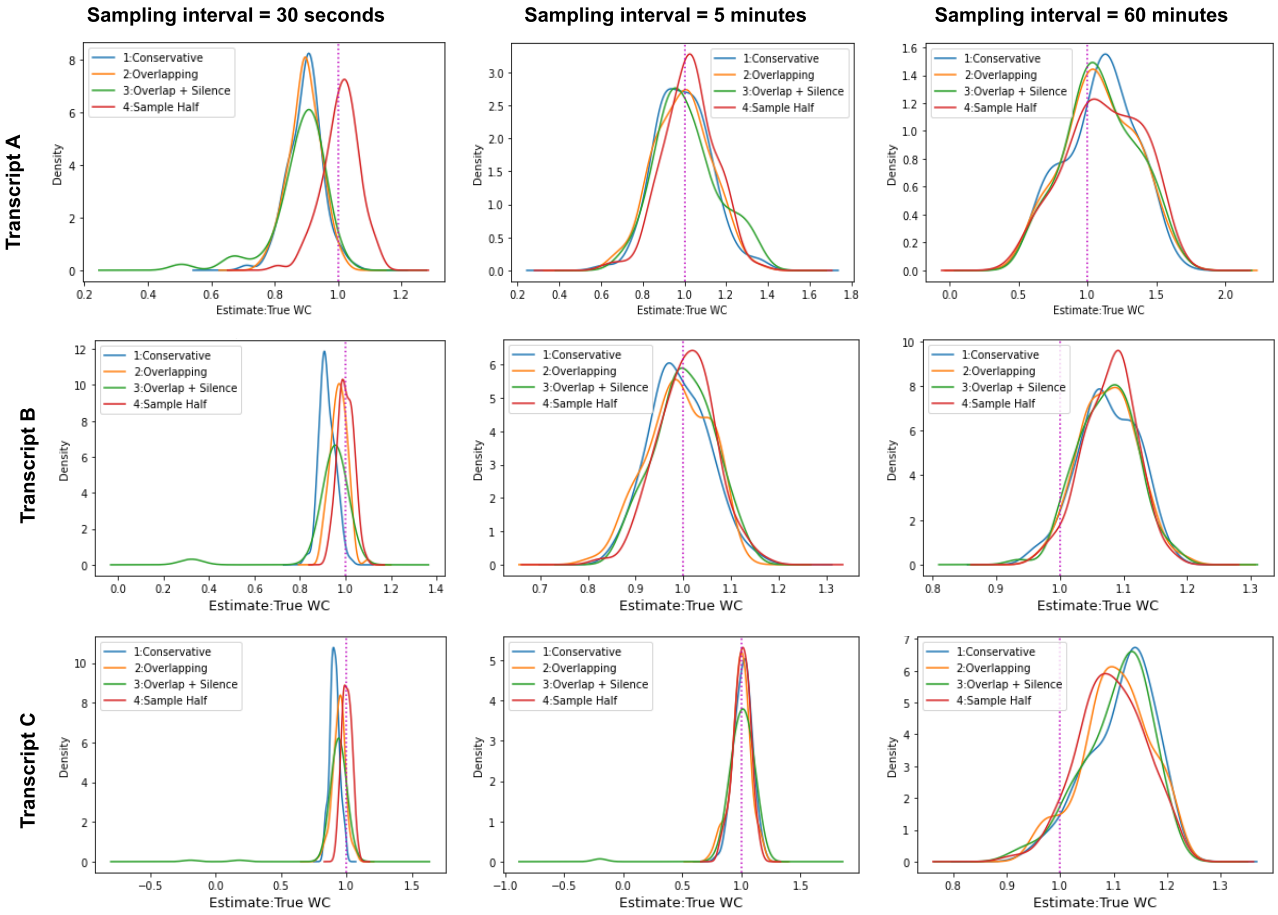

Figure 3: Kernel distribution estimation plots of the proportion of the estimated daylong word count to the true daylong word count of select sampling interval sizes (30 seconds, 5 minutes, 60 minutes) and a total sampled time of 120 minutes across the 4 implemented sampling methods (see Legend) over 100 simulations. A proportion less than 1.0 indicates that a given estimate is less than the true daylong count (underestimate) and a proportion greater than 1.0 indicates that a given estimate is greater than the true daylong count (overestimate). A dotted pink line at x = 1.0 is included for interpretability. The first row presents these plots for Transcript A, the second for Transcript B, and the third for Transcript C.

total sampled time, shorter sampling interval sizes produce more accurate estimates. Compared to sampling from a single continuous portion of the day, using shorter sampling intervals essentially allows sampling from the full range of the day, and thus produces more accurate estimates.

Additionally, these plots demonstrate that increasing the total sampled time produces more narrow distributions. Sampling more of the day sensibly will make estimates more precise; however, these plots reveal some caveats to this point. First, depending on the sampling interval size, sampling a full 2 hours may not be necessary. For example, in Transcript A, while the distributions across sampling interval sizes narrow when increasing total sampled time, these distributions are not markedly different from 100 minutes to 120 minutes, particularly when using shorter sampling intervals. This indicates that sampling 100 minutes rather than a full 120 minutes in conjunction with a shorter sampling interval may be sufficient and is an important point to note due to the time and resource intensive nature of manual transcription.

Further, regardless of total sampled time, larger sampling intervals (>10 minutes) present markedly different distributions across transcripts. In Transcript A (row 1 of Figure 4), estimate accuracy distributions for larger sampling intervals are wider but still generally centered around a proportion of 1.0. In Transcript B and even more so for Transcript C (row 2 and 3 of Figure 4, respectively), corresponding distributions are skewed to the right, indicating systematic overestimation with larger sampling intervals. The following analysis aims to further probe this disparity and provide an answer to why this systematic overestimation occurs for some, but not all, the transcripts.

Proportion of Speech Sampled

As discussed, for Transcripts B and particularly Transcript C, larger sampling interval sizes (> 10 minutes) produce consistent overestimates. Why is there systematic overestimation when employing larger sampling interval sizes for Transcripts B and C, but not for A, and how can this inform sampling choices?

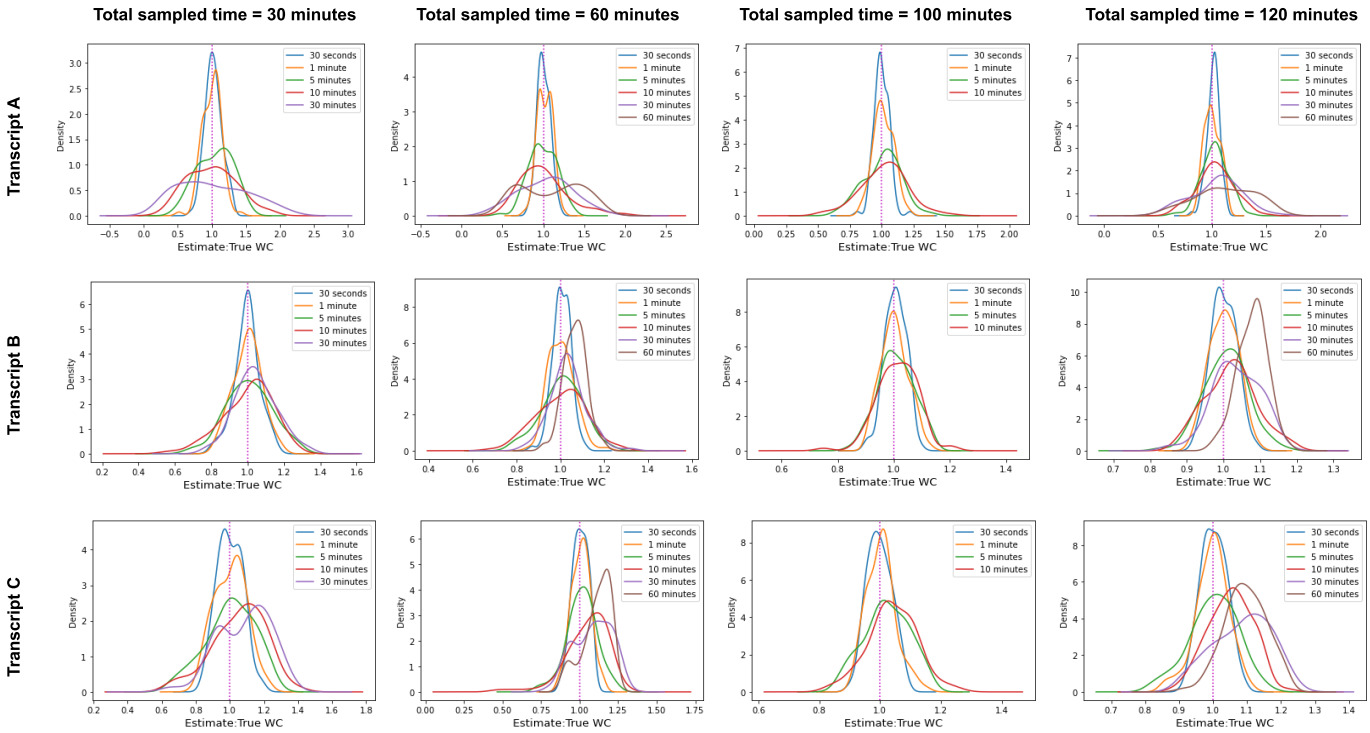

Figure 4: Kernel distribution estimation plots of the proportion of the estimated daylong word count to the true daylong word count for select total sampled times (30 minutes, 60 minutes, 100 minutes, 120 minutes) across different sampling interval sizes (30 seconds up to 60 minutes; see Legend) over 100 simulations. A dotted line at x = 1.0 is included for interpretability. The first row presents these results for Transcript A, the second for Transcript B, and the third for Transcript C.

A possible explanation may reside in the differences in actual speech distributions of the transcripts and its consequences on what and how much gets sampled. While all three transcripts came from daylong recordings of approximately 700-800 minutes, the transcripts' individual speech distributions are markedly different due to when the target child was in the home and/or when the family decided to record. Transcript A's speech is distributed more widely across the audio with about 63.8% of the recording containing transcribed speech. Conversely, Transcript B and C's speech is densely grouped in smaller consecutive periods of the audio, with about 45.8% and 23.9% of the recording containing transcribed speech, respectively. Thus, the proportion of speech sampled across the three transcripts differs notably from transcript to transcript and can potentially elucidate the disparity in the estimate distributions identified prior.

When sampling from Transcript A, interval sizes of 30 seconds totaling 120 minutes sample an average of 0.225 (SD = 0.013) of total words, whereas 60-minute intervals sample an average 0.250 (SD = 0.061) of total words. While both estimate accuracy distributions (see row 1 of Figure 4) are relatively centered around 1.0, the 30 second interval estimate distribution is much more narrow because sampling 120 minutes over 30 second intervals produces a smaller range of sampled word counts from which to extrapolate. Contrastingly, sampling with 60-minute intervals produces a wider range of total sampled words, and thus increases the likelihood of over- or under-estimating the daylong word count.

For Transcript B, while the proportion of speech sampled across simulations for both 30 second (M = 0.382, SD = 0.013) and 60-minute (M = 0.412, SD = 0.0156) sampling intervals have similar standard deviations, the 60-minute interval samples slightly more of the day. Transcript B contains an interval of speech in the middle of the day; however, its duration is too short to be sampled from with 60-minute intervals. This means that sampling is limited to the beginning and end of the day, which are denser intervals of input, and increases the likelihood of overestimation (row 2 of Figure 4). Likewise, Transcript C only contains speech in the beginning and end of the day, and the proportion of speech sampled with 30-second (M = 0.670, SD = 0.025132) versus 60-minute (M = 0.73, SD = 0.041027) intervals differs enough to generate relatively accurate and precise daylong estimates for smaller sampling intervals and systematic overestimates for larger sampling intervals.

Accuracies of Sampling Intervals

To further quantify accuracy rates across different sampling intervals, Figure 5 presents the number of estimates with a percent error greater than 10% using 30-second, 5-minutes, and 60-minute intervals for a total sampled time of 120 minutes.

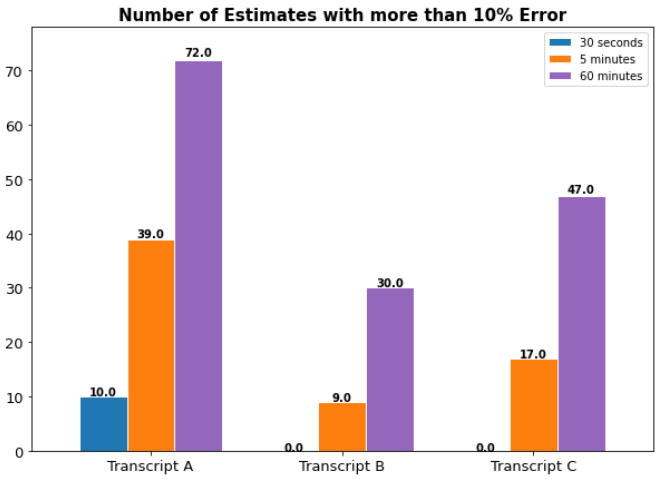

Figure 5: Number of estimates with more than 10% error using 30-second, 5-minute, and 60-minute sampling intervals and 120 minutes total sampled. For each sampling interval group, 100 estimates were obtained.

These results show how "lucky" an extrapolated estimate is when using smaller sampling intervals compared to larger sampling intervals. When using 60-minute intervals, 72 out of 100 estimates from Transcript A had a percent error greater than 10%; Transcripts B and C had corresponding counts of 30 and 47, respectively. Conversely, 30-second intervals were much more likely to produce more accurate estimates: in Transcript A, only 10 out of 100 estimates had a 10% error or greater; for Transcripts B and C, 0 estimates were observed at this measure. The notable differences in counts for large sampling intervals for Transcript A compared to Transcripts B and C is likely due to the differences in their respective speech distributions, as investigated in the previous section. Consistent with our previous analyses, these results demonstrate just how much more likely an estimate will be closer to the true total count when using smaller sampling interval sizes (30 seconds) compared to larger intervals (60 minutes).

Discussion

The present analyses explore the kinds of sampling choices one can make to optimize manual extrapolation with daylong audio data. Working from 3 fully transcribed recordings, we investigated sampling choices along 3 dimensions: (1) how overlapping utterances are handled (2) the total amount of time sampled, and (3) the size of individual sampling intervals. We demonstrated that the choice of sampling method in cases of overlapping utterances is most salient at smaller sampling interval sizes. Additionally, we quantified the effect of both total time sampled and sampling interval size. Smaller sampling intervals and longer durations of total sampled time lead to more accurate estimates, and we quantify how likely, in our audio transcripts, various sampling methods might be to yield accurate or extreme extrapolation estimates of features of interest.

While our results suggest that the smaller the sampling interval size the better, there are also practical limitations. Manually transcribing 30 second isolated intervals may not be efficient, because it involves meticulous records to extract, annotate, and reunite potentially hundreds of small audio clips. Thus using 1-minute or 5-minute intervals may yield a more manageable workflow. However, because past research has employed 30 second intervals (Ramírez-Esparza et al., 2014; 2017), and we see a substantial benefit for these short recordings we recommend it as a feasible and worthwhile sampling choice that should be thoughtfully implemented.

These results also demonstrate how increasing total sampled time sensibly improves estimate accuracy; however, in conjunction with smaller sampling intervals, the maximum analyzed total sampled time does not seem to notably improve estimate accuracy. Thus, with smaller sampling intervals, manually transcribing a full 2 hours may not be "worth it", and researchers can devote their finite resources and time elsewhere in their pipeline.

Finally, our analyses illustrate how the actual speech distribution of a given recording has downstream consequences for what and how much gets sampled and estimate accuracy. These analyses, in conjunction with our recommended step of identifying major intervals of silence (see Methods) underline the importance of preprocessing prior to sampling. While the true speech distribution of a recording certainly cannot be entirely known without full transcription, implementing these preprocessing steps will give a general sense of the recording's speech distribution; knowing whether speech is distributed more widely and variably (like Transcript A) or is densely grouped (like Transcripts B and C) help inform which sampling choices are better suited for the given recording. Recent efforts in developing automatic classifiers that intervals of sleep (Bang et al., 2021) present useful tools that can facilitate the preprocessing of this data and subsequently improve manual extrapolation.

To further elucidate the trade-offs between human effort and estimate accuracy, follow-up analyses are needed, particularly to thoroughly quantify how much total time sampled may be sufficient to ensure an accurate estimate. We also plan to include more daylong transcripts in our analyses that have differing speech distributions to validate our results. Additionally, we intend to simulate sampling with other linguistic features of interest, such as estimating select words, speech from select speakers, and child-direct/adult direct speech. Finally, we aim to compare manually extrapolated estimates with automated estimates from systems such as LENA to further inform daylong researchers about the methodological advantages and drawbacks of both alternatives.

Daylong audio presents an exciting avenue of research that has the potential to guide and enrich our theories of language development. As the establishment of thoroughly validated methods and tools is still in its early stages, we urge daylong researchers to thoughtfully implement their methodological choices.

References

Bang, J. Y., Kachergis, G., Weisleder, A., & Marchman, V. A. (2022). An Automated Classifier for Child-Directed Speech from LENA Recordings.

Bergelson, E., Amatuni, A., Dailey, S., Koorathota, S., & Tor, S. (2019). Day by day, hour by hour: Naturalistic language input to infants. *Developmental science*, *22*(1), e12715.

Casillas, M., Amatuni, A., Seidl, A., Soderstrom, M., Warlaumont, A., & Bergelson, E. (2017). What do Babies hear? Analyses of Child- and Adult-Directed Speech. In Proceedings of Interspeech 2017 (pp. 2093-2097). doi:10.21437/Interspeech.2017-1409.

Casillas, M., Brown, P., & Levinson, S. C. (2017). Casillas HomeBank Corpus. doi:10.21415/T51X12

Casillas, M., Brown, P., & Levinson, S. C. (2020). Early language experience in a Tseltal Mayan village. *Child Development*, *91*(5), 1819-1835.

Casillas, M., Brown, P., & Levinson, S. C. (2021). Early language experience in a Papuan community. *Journal of Child Language*, *48*(4), 792-814.

Clemens, L., & Kegel, C. (2021). Unique contribution of shared book reading on adult-child language interaction. Journal of Child Language, 48(2), 373-386. doi:10.1017/S0305000920000331

Cristia, A., Bulgarelli, F., & Bergelson, E. (2020). Accuracy of the language environment analysis system segmentation and metrics: A systematic review. *Journal of Speech, Language, and Hearing Research*, *63*(4), 1093-1105.

Cychosz, M., Cristia, A., Bergelson, E., Casillas, M., Baudet, G., Warlaumont, A. S., ... & Seidl, A. (2021). Vocal development in a large-scale crosslinguistic corpus. Developmental science, 24(5), e13090.

Fausey, C. M. & Mendoza, J. K. (2018). FauseyTrio HomeBank Corpus. doi: 10.21415/T5JM4R

Ferjan Ramírez, N., Hippe, D. S., & Kuhl, P. K. (2021). Comparing automatic and manual measures of parent–infant conversational turns: A word of caution. *Child Development*, *92*(2), 672-681.

Ferjan Ramírez, N., Hippe, D. S., Correa, L., Andert, J., & Baralt, M. (2022). Habla conmigo, daddy! Fathers' language input in North American bilingual Latinx families. Infancy : the official journal of the International Society on Infant Studies, 27(2), 301–323. https://doi.org/10.1111/infa.12450

Fields-Olivieri, M. A., & Cole, P. M. (2022). Toddler negative emotion expression and parent-toddler verbal conversation: Evidence from daylong recordings. Infant Behavior and Development, 67, 101711.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. Journal of child language, 35(3), 515-531.

Greenwood, C. R., Thiemann-Bourque, K., Walker, D., Buzhardt, J., & Gilkerson, J. (2011). Assessing children's home language environments using automatic speech recognition technology. Communication Disorders Quarterly, 32(2), 83–92. https://doi. org/10.1177/152574011036782

Hoff, E. (2006). How social contexts support and shape language development. Developmental review, 26(1), 55-88.

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & Hedges, L. V. (2010). Sources of variability in children's language growth. Cognitive psychology, 61(4), 343-365.

Lieven, E. (2016). Usage-based approaches to language development: Where do we go from here? Language and Cognition, 8(3), 346-368. doi:10.1017/langcog.2016.16

McDaniel, J., Yoder, P., Estes, A., & Rogers, S. J. (2020). Predicting expressive language from early vocalizations in young children with autism spectrum disorder: which vocal measure is best?. *Journal of Speech, Language, and Hearing Research*, *63*(5), 1509-1520.

Mendoza, J. K., & Fausey, C. M. (2021). Everyday music in infancy. *Developmental science*, *24*(6), e13122.

Montag, J. L. (2020). New insights from daylong audio transcripts of children's language environments. In CogSci.

Ramírez-Esparza, N., García-Sierra, A., & Kuhl, P. K. (2014). Look who's talking: Speech style and social context in language input to infants are linked to concurrent and future speech development. *Developmental science*, *17*(6), 880-891.

Ramírez-Esparza, N., García-Sierra, A., & Kuhl, P. K. (2017). The impact of early social interactions on later language development in Spanish–English bilingual infants. *Child development*, *88*(4), 1216-1234.

Räsänen, O., Seshadri, S., Lavechin, M., Cristia, A., & Casillas, M. (2021). ALICE: An open-source tool for automatic measurement of phoneme, syllable, and word counts from child-centered daylong recordings. *Behavior Research Methods*, *53*, 818-835.

Soderstrom, M., Casillas, M., Bergelson, E., Rosemberg, C., Alam, F., Warlaumont, A. S., & Bunce, J. (2021). Developing A Cross-Cultural Annotation System and MetaCorpus for Studying Infants' Real World Language Experience. Collabra: Psychology, 7(1), 23445.

Tamis-LeMonda, C. S., Kuchirko, Y., Luo, R., Escobar, K., & Bornstein, M. H. (2017). Power in methods: Language to infants in structured and naturalistic contexts. *Developmental science*, *20*(6), e12456.

VanDam M, Warlaumont AS, Bergelson E, Cristia A, Soderstrom M, De Palma P., MacWhinney B. HomeBank (2016) An online repository of daylong

child-centered audio recordings. Seminars in Speech & Lan., 37, 128-141.
VanDam, Mark (2018). VanDam Public Daylong HomeBank Corpus. doi:10.21415/T5388S
Warlaumont, A. S., Pretzer, G. M., Mendoza, S. & Walle, E. A. (2016). Warlaumont HomeBank Corpus. doi:10.21415/T54S3C
Warlaumont, A. S., Sobowale, K., & Fausey, C. M. (2022). Daylong mobile audio recordings reveal multitimescale dynamics in infants' vocal productions and auditory experiences. Current directions in psychological science, 31(1), 12-19.
Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. Psychological science, 24(11), 2143-2152.
Weiss, Y., Huber, E., Ferjan Ramírez, N., Corrigan, N. M., Yarnykh, V. L., & Kuhl, P. K. (2022). Language input in late infancy scaffolds emergent literacy skills and predicts reading related white matter development. Frontiers in human neuroscience, 16, 922552. https://doi.org/10.3389/fnhum.2022.922552
Xu, D., Yapanel, U., & Gray, S. (2009). Reliability of the LENA Language Environment Analysis System in young children's natural home environment. Boulder, CO: Lena Foundation, 1-16.