Classification errors distort findings in automated speech processing: examples and solutions from child-development research

Lucas Gautheron[1,2,*] Evan Kidd[3] Anton Malko[3]
Marvin Lavechin[4] Alejandrina Cristia[2]

Abstract

With the advent of wearable recorders, scientists are increasingly turning to automated methods of analysis of audio and video data in order to measure children's experience, behavior, and outcomes, with a sizable literature employing long-form audio-recordings to study language acquisition. While numerous articles report on the accuracy and reliability of the most popular automated classifiers, less has been written on the downstream effects of classification errors on measurements and statistical inferences (e.g., the estimate of correlations and effect sizes in regressions). This paper proposes a Bayesian approach to study the effects of algorithmic errors on key scientific questions, including the effect of siblings on children's language experience and the association between children's production and their input. In both the most commonly used Language ENvironment Analysis (LENA™), and an open-source alternative (the Voice Type Classifier from the ACLEW system), we find that classification errors can significantly distort estimates. For instance, automated annotations underestimated the negative effect of siblings on adult input by 20–80%, potentially placing it below statistical significance thresholds. We further show that a Bayesian calibration approach for recovering unbiased estimates of effect sizes can be effective and insightful, but does not provide a fool-proof solution. Both the issue reported and our solution may apply to any classifier involving event detection and classification with non-zero error rates.

Keywords: language acquisition, long-form recordings, speech processing, classification bias, event detection, latent variable modeling

[1]University of Wuppertal, Germany

[2]Laboratoire de Sciences Cognitives et Psycholinguistique, Département d'études cognitives, ENS, EHESS, CNRS, PSL University, Paris, France

[3]School of Literature, Languages and Linguistics, Australian National University, Canberra, Australia

[4]Computational Psycholinguistics Lab, Massachusetts Institute of Technology, United States

Contents

1 Introduction

Children's behavior and their environments are increasingly described through automated analysis of data collected from wearables. Pioneers in such techniques, researchers working on language acquisition have shown the promise of using automated classifiers to analyze long-form audio recordings, thus enabling the processing of naturalistic data at unprecedented scale (Bergelson et al., 2023). A common application in this context is the automatic segmentation and classification of speech into voice categories for measuring speech afforded to and produced by infants. A growing number of studies document and discuss the accuracy of these automated classifiers (such as Language ENvironment Analysis (LENA™)), by comparing human and machine annotations of the same audio clips (Xu et al., 2009; Lavechin et al., 2025). Concern has been raised about unexpected cases of low recall and precision (e.g., a precision of 27% for the recognition of the child wearing the device in Gilkerson et al. 2015), as well as trends for confusion across speaker types (Lehet et al., 2021). To our knowledge, less attention has been paid to how classification errors propagate through subsequent analyses. This paper examines a critical question: To what extent do errors in automated speech processing systems like LENA™ affect downstream measurements and scientific conclusions? Specifically, we investigate how speaker tagging misclassifications (e.g. confusing child speech with adult speech) impact measurements of children's linguistic input and production, as well as the effect sizes based on these measurements. Our work reveals significant issues across multiple types of measurements. For instance, we find that the negative effect of siblings on the quantity of adult input children receive can be underestimated by 20–80% (and potentially declared statistically insignificant) when algorithmic errors are not taken into consideration. In order to address this issue, we introduce a Bayesian approach for recovering unbiased measurements correcting for algorithmic errors. The methodological insights gained here may also apply to signals captured by other wearable technologies (e.g., video; Long et al. 2024); and generally, to any case where event detection is performed in conjunction with classification using machine learning algorithms.

1.1 The case of voice type classifiers when describing early language acquisition

Day-long audio recordings of children's language experience collected with wearable devices have become widespread thanks to their richness and ecological validity. The adoption of this methodology has been facilitated by LENA™, a user-friendly commercial solution simplifying data collection and analysis. LENA™ provides both the recording device and the software for automatically annotating the audio. The latter is important, given that this technique produces a large amount of data (up to 16-24 hours of audio per

session in the case of LENA™, to be multiplied by the number of children and the number of sessions per child), which would be impossible to analyze entirely by hand. Among other things, LENA™ includes a diarization algorithm that detects speech (*event detection*) and attributes it to one of four different types of speaker (*classification*): the child wearing the recording (CHI), another child – e.g. a sibling – (OCH), female adult (FEM) and male adult (MAL).[1] With additional processing steps that build on this essential diarization algorithm, LENA™ provides a number of metrics, including the child vocalization count (CVC; i.e., the number of speech-like segments attributed to the key child). Researchers were quick to consider potential errors in the algorithm. Processing audio data collected directly from children in noisy real-life conditions is challenging, and the metrics returned by LENA™ (or any other classifier) are far from perfect. Yet, LENA™ is generally considered to have been sufficiently validated, with a recent meta-analysis (Cristia et al., 2020) finding that the correlation between human and LENA™ CVC in the same audio clips averaged r=.77 (N=5). Moreover, another meta-analysis found that CVC correlated with concurrent and/or longitudinal standardized measures of language (r=.33, N=10; Wang et al. 2020).

While validation studies are undeniably important, we believe that computing performance metrics is not sufficient and that there has not been enough consideration about how algorithms' errors may impact conclusions. We illustrate this on the widely used vocalization count measure. Figure 1 shows the segmentation into speaker categories of a 30s audio clip made by a human expert, the proprietary LENA™ algorithm, and its open-source alternative, Voice Type Classifier (VTC) (Gilkerson et al., 2015; Lavechin et al., 2020). Errors in the automated annotations result in erroneous vocalization counts. For instance, in this example, both LENA™ and VTC incorrectly report two vocalizations from siblings.

In the literature, estimated vocalization counts are often directly plugged into statistical models to perform measurements addressing a variety of questions. This ignores the fact that the quantity of speech attributed to a given voice type can be affected by the quantity of speech from others (Figure 2). For instance, the proportion of female adult speech can be overestimated due to children being confused with a female adult, or distorted due to male and female adults being confused with one another (2a). Classification errors can have effects not only on our estimates of these quantities, but also on our estimates of the association strength between these quantities or with other variables. From a causal inference perspective, classification errors open "biasing paths"[2] that can cre-

[1] LENA™ also returns other classes, such as TV/electronic noise. Since these have been more seldom the target of methodological work and less commonly used in scientific research by and large, we do not consider them here.

[2] It is useful to frame this problem in terms of causal diagrams (as Directed Acyclic Graphs, or DAGs), in which causal relationships between variables are represented by directed arrows. Causal paths introduce correlations between variables: if, e.g., exposure $[e] \rightarrow$ mediator $[m] \rightarrow$ outcome $[o]$ – that is, an exposure causes a mediator which causes an outcome –, e, m and o will all appear to be correlated.

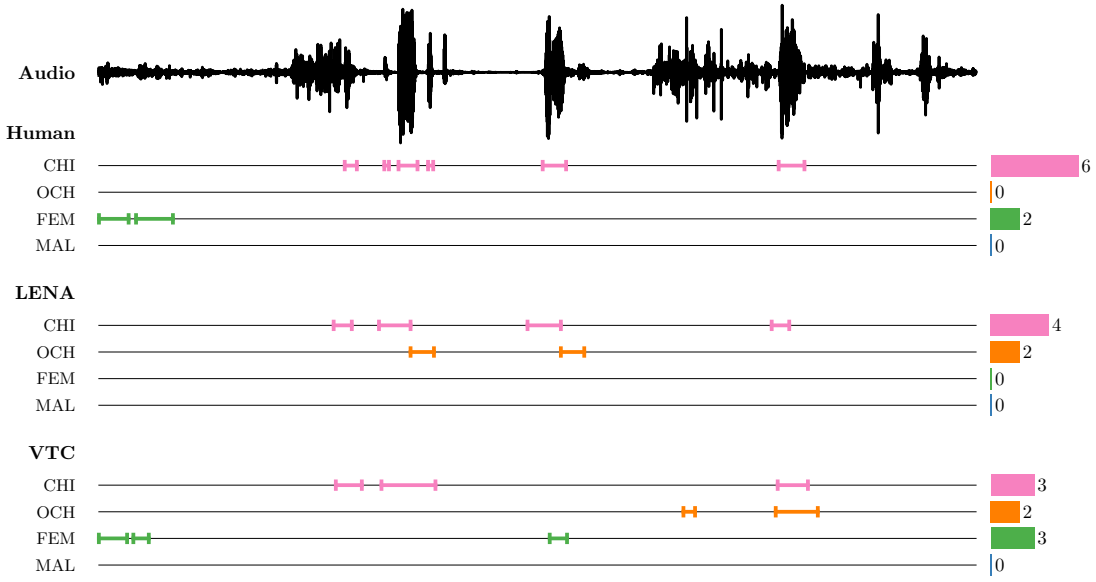


Figure 1: 30-second sample of a daylong recording annotated by a human expert and two algorithms: LENA™, and VTC. CHI refers to the child wearing the recording device; OCH refers to other children; FEM and MAL refer to female and male adults. A segment of speech is referred to as a "vocalization" (for instance, the expert found two female adult vocalizations in this portion of audio, but LENA™ found none). Vocalization counts are shown to the right.

ate completely spurious correlations between variables in sometimes unpredictable ways. This potentially leads us to over-/under-estimate the effect of an association; to believe that an association between two variables exists when it does not; or worse, to reach incorrect conclusions about the *sign* of an effect. For instance, in Figure 2b, female adult speech incorrectly labeled as child speech may lead to spurious correlations between our measurements of adult speech (input) and child speech (output). In Figure 2c, the effect of siblings on adult input may be similarly distorted by classification errors, which open up a completely spurious biasing path between "siblings" and female/male adult speech.

While the latter issues affect algorithms in isolation, another problem that has been overlooked is the comparison of measurements extracted from different algorithms (e.g. LENA™ versus VTC). For instance, LENA™ reports much less speech than VTC, especially for certain speakers – as an example, our data contains a recording where LENA™ finds less than 200 vocalizations from male adult(s) and VTC finds more than 2000. This

Causal inference aims to infer causal relationships (i.e. causal paths) given correlation patterns relating multiple variables. Per Textor and Liskiewicz (2012), "any type of bias that can be expressed in the formal framework of causal diagrams corresponds to [...] a *biasing path*." . More precisely, "[relevant] causal paths start at the exposure [e], contain only arrows pointing away from the exposure [e], and end at the outcome [o]. That is, they have the form $e \rightarrow x_1 \rightarrow \cdots \rightarrow x_k \rightarrow o$. Biasing paths are all other paths from exposure to outcome [e.g.] $e \leftarrow x_1 \rightarrow \ldots x_k \rightarrow o$" (Textor, 2015). In the latter case, x_1 is typically called a *confounder*.

is concerning because researchers typically analyze their data using only one algorithm, making it unclear how algorithm choice affects their scientific conclusions.

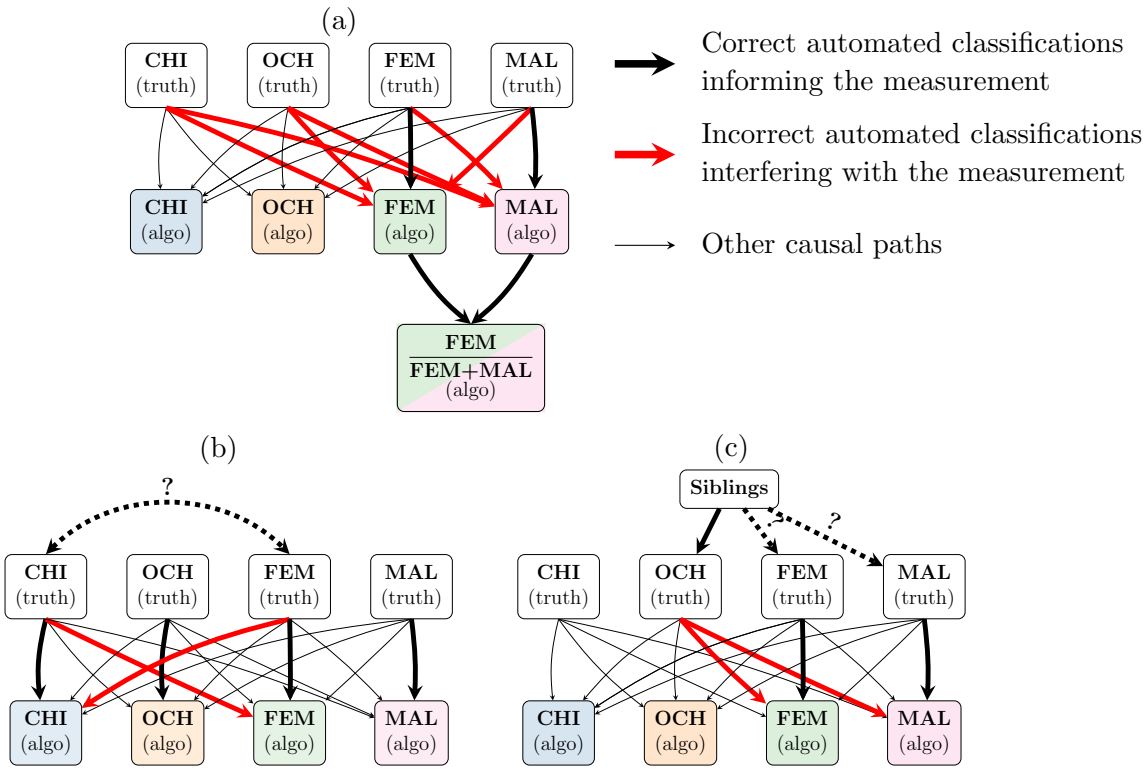


Figure 2: The quantity of speech attributed to each speaker ("CHI", "OCH", "FEM", "MAL", i.e. the key child, other children, female adults, and male adults) in each recording by an algorithm only indirectly reflect the true quantities. In reality, speaker classification errors can distort measurements and create spurious correlations in the quantities of speech attributed to each speaker. **(2a) Measurements of speech quantities.** The nature of the input to children may be misrepresented as a result of classification errors. For instance, the proportion of female adult speech can be distorted due to incorrect inferences about the speaker's type and gender. **(2b) Associations between speakers.** An increase in female adult speech may trigger an increase in detected amounts of both female adult (black arrow) *and* child speech (red arrow), and vice-versa, creating the appearance of an association between the two speakers. **(2c) Effect of independent variables on speech quantities.** Spurious associations can also bias inferences about the effect of independent variables on speech behavior. For example, we might draw incorrect conclusions about the existence and direction of an effect of siblings on the quantity of speech received from adults (dashed lines) if speech from siblings is incorrectly classified as adult speech (then, children with siblings might falsely appear to receive more input from adults).

1.2 Previous relevant work

Thus, while the accuracy, reliability, and validity of LENA™ has been well documented in previous works for many languages (for meta-analyses see Cristia et al., 2021; Wang

et al., 2020; see also Bastianello et al., 2023; Bruyneel et al., 2021; McDonald et al., 2021; Cristia et al., 2024), the downstream effect of confounding bias due to classification errors on correlational analyses has not been assessed. This is necessary given that two measures can be individually reasonably accurate (i.e., strongly correlated with human judgment), reliable (i.e., stable across repeated measurements), and valid (i.e., strongly correlated with independent relevant metrics), and nevertheless spuriously correlate to each other. Thus, the most commonplace strategies for evaluating measurements in psychology are ineffective against classification bias. In the context of long-form recordings for language acquisition, the issue was previously raised in Cristia et al. (2023), a descriptive report on 38 children's language input and output, and in which an attempt was made to discern whether correlations observed between speakers' speech quantities were entirely consistent with classification errors. The present work addresses this concern more generally by covering a wider range of research questions and proposing a solution for recovering unbiased statistical estimates.

In parallel, the incidence of bias resulting from prediction errors has become more widely recognized in other contexts, due to the increasing recourse to machine learning for data processing. Most prominently, Angelopoulos et al. (2023b) proposed a general approach to the issue (prediction-powered inference), which they have demonstrated on an array of datasets from different fields (biology, astrophysics, ecology, etc.). However, their strategy is currently limited to simple usages (e.g. ordinary least-squares linear-regression) and lacks flexibility for complex hierarchical models such as those examined in the present paper[3]. In addition, their approach requires ground-true values/labels for a substantial amount of observations. This makes it hard to transpose to our case, since no recording could be entirely annotated by hand (only a handful of clips at best). In a different context, TeBlunthuis et al. (2024) insisted that "current practices of 'validating' [automated classifiers] by making misclassification rates transparent via metrics such as the F1 score [...] provide little safeguard against misclassification bias". Independently from us and building upon prior literature on measurement errors (Carroll et al., 2006, Ch. 8, 15), they proposed a solution ("Maximum Likelihood Adjustment") conceptually similar to ours but limited to classification tasks[4]. In a nutshell, they propose to average out statistical estimates over all possible true labels, weighted by their respective likeli-

[3]Their approach relies on a *rectifier*, i.e, a function that compensates for measurement errors in the estimator of a quantity of interest (e.g., a linear or logistic regression coefficient) by leveraging manual annotations. Their strategy can in principle be applied to a broad class of problems (in particular, parameter estimation via convex or non-convex loss minimization problems of a certain form). Interestingly, no assumption about the underlying machine-learning classifier is required. However, deriving the rectifier is not a trivial task for an arbitrarily complex model. For very simple models, ready-to-use implementations have become available (Angelopoulos et al., 2023a; Salerno et al., 2025). Efforts to enhance the flexibility of this line of approach are ongoing (Miao & Lu, 2024).

[4]The authors present their method as an improvement over previous approaches such as Multiple Imputation (MI). In MI, multiple analyses are performed and pooled together given different imputations of missing or noisy data (Blackwell et al., 2017).

hood, given the labels predicted by a classifier and possibly other covariates. TeBlunthuis et al. (2024) set aside a Bayesian treatment for future work, acknowledging that [this] "may provide additional strengths in flexibility and uncertainty quantification". Therefore, the current paper makes a distinct contribution with respect to TeBlunthuis et al. (2024), through a flexible Bayesian approach and by overcoming the restriction to classification tasks. Additionally, we are motivated to make these contributions accessible to a wider readership interested in behavioral research methods.

1.3 Present work

The present paper evaluates the effect of speaker classification errors in two popular algorithms on the estimates of quantities and effects relevant to language acquisition. These algorithms are LENA™, the historically dominant solution in the field, and VTC, a state-of-the-art alternative gaining traction (Laudańska et al., 2025). First, using annotations from VTC and LENA™ annotations of the very same longitudinal audio data (1 400 recordings of ∼8 hours each from 237 children across six corpora; see Section §2.3), we demonstrate that correlations between different types of speakers in these two algorithms are distorted in a way that is consistent with what is expected from classification errors. We then assess the consequences of these distortions for downstream measurements relevant to language acquisition research, according to three categories (Figure 2): (a) direct measurement of speech quantities (in our example, the proportion of female adult input); (b) measurements of associations between speech quantities (the short and long term effects of input on output); (c) measurements of the effect of an independent variable on speech quantities (respectively, the effect of the child's age on its vocal production, and the effect of sibling number on the child's experience).

To this end, we fit a multi-level hierarchical model that simultaneously models the relationship between the *true* but unobserved vocalization counts of each different talker type (male adults, female adults, other children, and the key child) and the vocalization counts provided by diarization algorithms. This allows the model to estimate to what extent the *true* vocalization counts are due to developmental effects (with more input leading children to vocalize more over time) as well as household effects (specifically, the presence of siblings leading to increases in vocalization counts attributed to other children).

As a brief summary, we find that classification bias can significantly distort estimates of speech quantities and effect sizes in downstream analyses, sometimes even leading to incorrect conclusions about the existence, size, and (possibly) the direction of a correlation/effect. Furthermore, the impact of classification errors varies depending on the precise effects being estimated. For instance, while the effect of having siblings on the quantity of speech from each speaker is significantly distorted by measurement errors,

distortions are smaller for developmental effects unfolding throughout child development, such as the long-term effect of input on output. This paper provides a Bayesian calibration model for producing unbiased estimates of correlations and effect sizes using algorithmic vocalization data, which may be most useful to technically proficient readers. Our proposed model combines automated annotations together with human annotations with adequate weighing, an approach that improves the agreement between LENA™ and VTC in all cases except one: the association between adult input and child output. We speculate that this association may be obscured by algorithmic distortions of the temporal dynamics of adult-child interactions at short-time scales unaccounted for in our model (e.g., the inability of LENA™ to support overlapping speech). Finally, we also provide a Python package enabling scientists to simulate the impact of classification errors on their own analysis (Gautheron, 2025).

1.3.1 Outline of our methodological approach

The following is a high-level description of our methodological approach, with details being provided in the Methods section. We start by specifying a model of speech behavior that embeds our theoretical assumptions and/or the hypotheses we are willing to evaluate. In broad terms, our model of speech behavior specifies a small set of factors that may predict children's speech input and output. Children's input composition may vary as a function of how many siblings the child has; and the key child's output may vary as a function of their age, input, and random individual variation. We do not mean to imply that these are the only factors researchers should care about – we selected them primarily to illustrate a range of research questions and evaluate the effect of classification errors on reasonably motivated analyses. Conceptually, individual researchers will assume a model like the one we posited in order to answer a research question, e.g. whether the number of siblings explains significant variance in how many adult vocalizations infants are afforded. Their models of speech behavior may thus vary from our own.

Next, we embed this model of speech behavior in a larger model that also take into account the fact that we do not observe children's "REAL" number of adult vocalizations, but instead estimate them through an algorithm. To this end, we include the effect of the algorithm in the data generating process, treating the true vocalization counts as latent (i.e. unobserved) variables. Fortunately for the current research, several datasets have been partially annotated by humans, which means that (for very small extracts out of the long-form recordings), we have gold standard estimates of how many vocalizations were uttered by the key child, other children, as well as male and female adults. For this initial foray in studying how algorithmic errors may affect scientific conclusions, we assume that the most relevant features of the algorithm relate to the model's tendency to miss vocalizations, assign them to the wrong speakers, as well as incorrectly break up

or lump vocalizations.

Using a Bayesian approach, then, our task becomes estimating each of the parameters in our full model, using snippets of audio for which we have both algorithmic and human annotations to estimate parameters related to the model of the algorithm, and using all data (annotated by humans or algorithms) to inform parameters related to our model of speech behavior. Our approach is highly flexible, since it decouples the statistical model of speech behavior from the model of the algorithm's behavior: both can be refined in parallel as two distinct modules, regardless of their respective complexity. Finally, in contrast to TeBlunthuis et al. 2024, the proposed strategy goes beyond pure classification tasks by considering also errors emerging from the detection and diarization/segmentation aspects. This was crucial in our case, since the algorithms considered in this paper make a variety of errors beyond the misidentification of speakers (e.g., they can also fail to detect certain vocalizations, or fragment single vocalizations into multiple ones).

The technical details of our approach are elaborated in Section (§2). Readers mostly interested in our findings and their implications may jump to the Results section directly (§3).

2 Method

In this Section, we introduce the assumed model of speech behavior (Section §2.1), the Bayesian calibration approach (Section §2.2), and finally the data (Section §2.3).

2.1 Model of speech behavior

Our multi-hierarchical model implements several assumptions (see Figure 3). First, the quantity of vocalizations by each speaker class (CHI, OCH, FEM, and MAL) is thought to potentially vary across children. Second, the number of siblings a child has may affect speech quantities by OCH and ADU (i.e., FEM and MAL), but not CHI directly. Third, we also assume a random child-specific effect of development on children's speech output. Specifically, our model assumes that children's speech quantities at birth are equivalent (i.e. individual newborns do not differ from each other), with random individual variation emerging as children age (using a Generalized Linear Model with a log link function). Finally, the model also assumes a long-term effect of adult input on children's output (i.e., an effect of adult speech at a child-level that interacts with the children's age). All effects of age are assumed to be log-linear, up until a threshold (24 months) after which they plateau (this threshold was validated via a change-point model). The precise model specification, including the priors on every parameter, is described in Section §A.1.1. In addition, in Section §3.1, we consider a simplified version of this model, in which correlations between speakers at the recording level and at the child-level are implemented

with multivariate lognormal distributions (see Appendix §A.1.2).

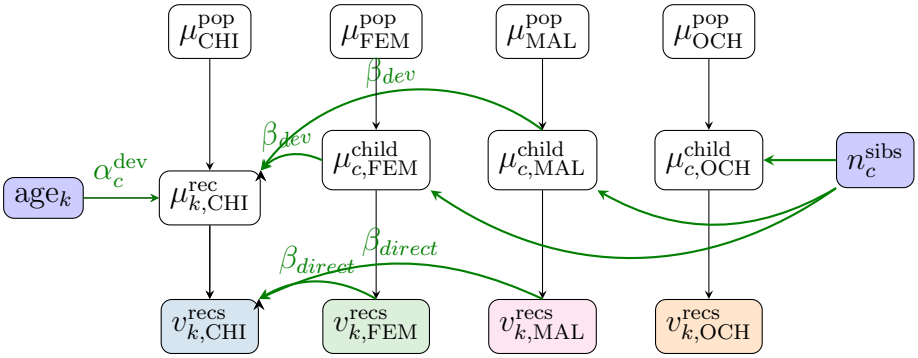


Figure 3: Model of speech behavior. Observed variables (vocalization counts for each speaker and recording, child age, and siblings number) are shown in blue, latent variables in white. Indices k designate recordings, and c designates a child.v^{recs}_k is the vocalization count of each speaker class in each recording. Variables μ represent the expected speech rates per speaker at each level (population, corpus, and child). α^{dev}_c is the random effect of age on the children's output (which is assumed to be distributed around a mean value α_{dev}). It is also assumed that the expected quantity of adult speech at the child level has a long-term effect on children's speech (β^{dev}), which interacts with age.

2.2 Bayesian calibration of algorithmic vocalization counts

2.2.1 Principle

Given the above model of speech behavior (section 2.1), we seek to estimate the probability distribution of various parameters $P(\theta|D_{meas})$, given the measured vocalization data D_{meas}. However, if D_{meas} is a noisy and biased estimate of reality (as is the case for vocalization counts derived from diarization algorithms rather than the true quantities), treating it as truth (as shown in Figure 2) can lead to incorrect conclusions. We therefore propose a Bayesian calibration approach to deal with observations that are noisy and biased due to classification errors.

The principle is to infer the probability distribution of the unobserved *true* amounts of vocalizations of each speaker ($\hat{D}$) given the algorithm output (D_{meas}), and to plug those estimates into the model of behavior rather than D_{meas} directly. If θ are the parameters of interest of the model, D_{meas} are the data produced by the algorithm, $\hat{D}$ are the true but unobserved quantities, and ν are nuisance parameters that describe algorithm behavior, then the posterior distribution of θ (the parameters of interest) given the observed data D_{meas} is:

$$P(\theta|D_{meas}) \propto \int \underbrace{P(\hat{D}|\theta)P(\theta)}_{\text{speech behavior}} \underbrace{P(D_{meas}|\hat{D},\nu)P(\nu)}_{\text{algorithm behavior}} d\hat{D}d\nu \quad (1)$$

The posterior distribution is factored into two terms. The first term encodes the

assumed model of speech behavior (described in Section §2.1 and represented in Figure 3). The second term encodes the algorithm's behavior, in terms of the probability that the algorithm attributes certain quantities of vocalizations per speaker (D_{meas}) given the true unobserved quantities $\hat{D}$ (if the algorithm was perfect, we would have $P(D_{\text{meas}}|\hat{D}) = 1$ if and only if $D_{\text{meas}} = \hat{D}$, and 0 otherwise). The nuisance parameters ν characterize the algorithm's behavior (typically, the classifier's confusion matrix), and are a priori unknown. To learn ν, one must add calibration data, for which both the algorithm's output (D_{calib}) and the ground truth ($\hat{D}_{\text{calib}}$) are observed. This gives:

$$P(\theta|D_{\text{meas}}, D_{\text{calib}}, \hat{D}_{\text{calib}}) \propto \int \underbrace{P(\hat{D}|\theta)P(\theta)}_{\text{speech behavior}} \underbrace{P(D_{\text{meas}}|\hat{D}, \nu)P(\nu|D_{\text{calib}}, \hat{D}_{\text{calib}})}_{\text{algorithm behavior}} d\hat{D}d\nu \quad (2)$$

To the extent that the model of the algorithm's behavior is valid *and* the "ground-truth" data is correct, the posterior distribution of θ retrieved by computing this integral will be unbiased. However, the uncertainty induced by the behavior of the algorithm will widen the posterior distribution of θ, given that different values of $\hat{D}$ (the true quantities of speech) are compatible with the algorithm output D_{meas}; in other words, this approach trades bias for variance.

Figure 4 represents the full model which combines the model of speech behavior and the model of the algorithm behavior. This figure shows how the algorithmically derived vocalization counts are no longer plugged directly into the model. It also shows how calibration data can be used to learn the relationship between the true and the measured vocalization counts. This approach is highly flexible, given that the model of speech behavior and the model of the algorithm can be elaborated and improved separately as independent modules (in contrast to Angelopoulos et al., 2023b).

2.2.2 Model of the algorithm's behavior

The calibration approach requires a model of the algorithm in terms of the probability that it outputs certain values given the unobserved true amount of vocalizations of each speaker class $i \in \{\text{CHI}, \text{OCH}, \text{FEM}, \text{MAL}\}$, which will be further referred to as v_i. We assume that each of the vocalizations from each speaker i causes the algorithm to attribute a random amount (0, 1, 2, ...) of vocalizations to each speaker j, resulting in a total of n_{ij} vocalizations attributed to j as a result of the true vocalizations from i. The only observable quantity is, in fact, $n_j = \sum_i n_{ij}$, the total amount of vocalizations attributed to each speaker j by the algorithm. Different assumptions could be made about how n_{ij} is generated, given v_i, the unobserved true amount of vocalization from each speaker i. For instance, one can assume a binomial process, such that the v_i vocalizations are detected and attributed to j with probability λ_{ij} (that is, $n_{ij} \sim \text{Binomial}(v_i, \lambda_{ij})$). However, some

vocalizations are detected as not one but multiple vocalizations (the algorithm breaks them down into multiple segments), which a binomial process would fail to capture. We therefore consider a generalized Poisson distribution (from Efron 1986), such that $n_{ij} \sim \text{DPO}(\lambda_{ij} v_i, \tau)$ with mean $\lambda_{ij} v_i$ and variance $\lambda_{ij} v_i / \tau$[5].

In this model, (λ_{ij}) is the confusion matrix of the algorithm; the diagonal (λ_{ii}) measures the rate of true positives, and the non-diagonal elements $\lambda_{i,j\neq i}$ measure the rate of false positives due speaker misidentification. The confusion rates λ_{ij} are assumed to vary from one recording to another (due to unpredictable variations in recording conditions for instance), such that (λ_{ij}^k) (the confusion rates for a particular recording k) are drawn from Gamma distributions with means μ_{ij} and shapes $\alpha_{ij} \sim \text{Pareto}(1, 1.5)$ (truncated to values ≥ 1[6]). λ, μ, α, and τ are nuisance parameters that can be partially learned from calibration data.

Ultimately, we implement the model in Stan (Carpenter et al., 2017), which uses a variant of Hamiltonian Monte-Carlo and therefore cannot directly sample latent discrete parameters such as n_{kij} and n_{kj}[7]. In the case of the calibration data, for which the ground truth (v_{ki}) is known, the solution is to marginalize over the discrete latent parameters n_{kij}, which comes down to computing the sum (3):

$$P(n_{kj} = n | v_{k1}, \ldots, v_{kC}) = \sum_{0 \leq n_{kij} \leq n} \prod_{i=1}^{C} P(n_{kij} | v_{ki}) \cdot \delta(n - \sum_{i=1}^{C} n_{kij}) \tag{3}$$

The joint knowledge of the algorithm's vocalization counts (n_{kj}), and the true counts (v_{ki}) in manually annotated clips of audio allows us to learn the distribution of confusion rates (λ) across recordings. There is one computational caveat: the amount of combinations to be summed over becomes combinatorially large for large values of v_i. Therefore, manually annotated clips are split into windows of 15s, which keeps v_i reasonably small with each window (originally, the duration of manually annotated clips varied between 15s and 5 minutes, always being a multiple of 15s). In fact, larger temporal windows would be less informative. However, we do not think it wise to use even shorter time windows, since it would introduce boundary effects, and we found 15s to be a good compromise between tractability, information-value, and bias.

For most of the audio, only automated annotations are available, and the true values v_{ki} are unknown: they are latent parameters, and the goal is precisely to measure their distribution. To this end, we approximate v_{ki} as a continuous parameter – which Stan

[5]We assume $\tau \sim$ Exponential(1), such that the model can accomodate both under- and over-dispersion. We use the Stan implement of the Double Poisson distribution proposed in Pustejovsky (2024). We find $\tau \sim 1.4$ for VTC and $\tau \sim 1.8$ for LENA™, corresponding to underdispersion.

[6]α is difficult to identify, especially for low values of μ. Ideally, we would need more extensive human annotations within each manually annotated recording. In this work, we limited ourselves to extant annotations.

[7]See https://mc-stan.org/docs/stan-users-guide/latent-discrete.html.

can conveniently sample from – and assume that:

$$n_{kj} \sim \text{DPO}(\sum_{i=1}^{C} \lambda_{kij} v_{ki}, \tau) \tag{4}$$

where DPO designates the generalized Poisson distribution from Efron (1986).

As with any model, the above simplifies reality in numerous ways. First, the effect of true vocalizations from each speaker is assumed to add linearly, which might not be a good approximation in case of overlapping speech. This could slightly bias estimates of correlations between input and output if those were driven by dense interactions over short time scales. In fact, the model ignores the fact that LENA™ does not handle overlap between speakers. Second, confusion rates are assumed to be random and independent from variables such as the child's age, or population level effects (e.g. differences in environment or language). Yet, if, say, an algorithm detects vocalizations from older children more accurately, this could lead to overestimating the increase in output over time. Our approach (just like that of TeBlunthuis et al. 2024) can in principle account for such biases. However, we found that we did not have access to enough human annotations to incorporate these effects directly into our full model, due to poor identification. Nevertheless, we independently assessed the effect of the child's age and the environment (urban vs rural) in Appendix §A.3, which did not reveal unambiguous and significant effects (see also Peurey et al. 2025). We recommend that additional work reflects on the potential consequences of confusion rate depending on these factors. Finally, our approach relies on the assumption that human annotations provide a reliable ground truth, but humans make mistakes. Agreement between human annotators in typical longform recordings has been measured in terms of the average F-score[8] (across speaker types), yielding $F_1 \simeq 0.70$, which is far from perfect (Kunze et al., 2025). By contrast, VTC achieves $F_1 \simeq 0.51$[9].

For purposes of validation, we evaluated whether the model was able to predict the output of the algorithms, given the true amount of vocalization from each speaker in manually annotated clips (Appendix §A.2.1). We found that the model is generally able to anticipate the output of the algorithms given the human annotations in the calibration data. In addition, the model correctly identifies the confusion matrix in simulated annotations (Appendix §A.2.2). However, it does not produce very confident predictions; this is because the behavior of the algorithms is highly stochastic (i.e., confusion rates may vary a lot across recordings). We also found that the calibration procedure improved the

[8]The F-score is the harmonic mean of precision ($\frac{\text{true positives}}{\text{true positives+false positives}}$) and recall ($\frac{\text{true positives}}{\text{true positives+false negatives}}$).

[9]These results correspond to frame-level accuracy metrics, which depend also on the annotators' ability to precisely locate the onset/offset of a vocalization. This is a more difficult task compared to merely identifying vocalizations counts for each speaker (regardless of their precise endpoints) which is all that is needed in our case. Therefore, we expect that human annotations provide sufficient ground truth for our case.

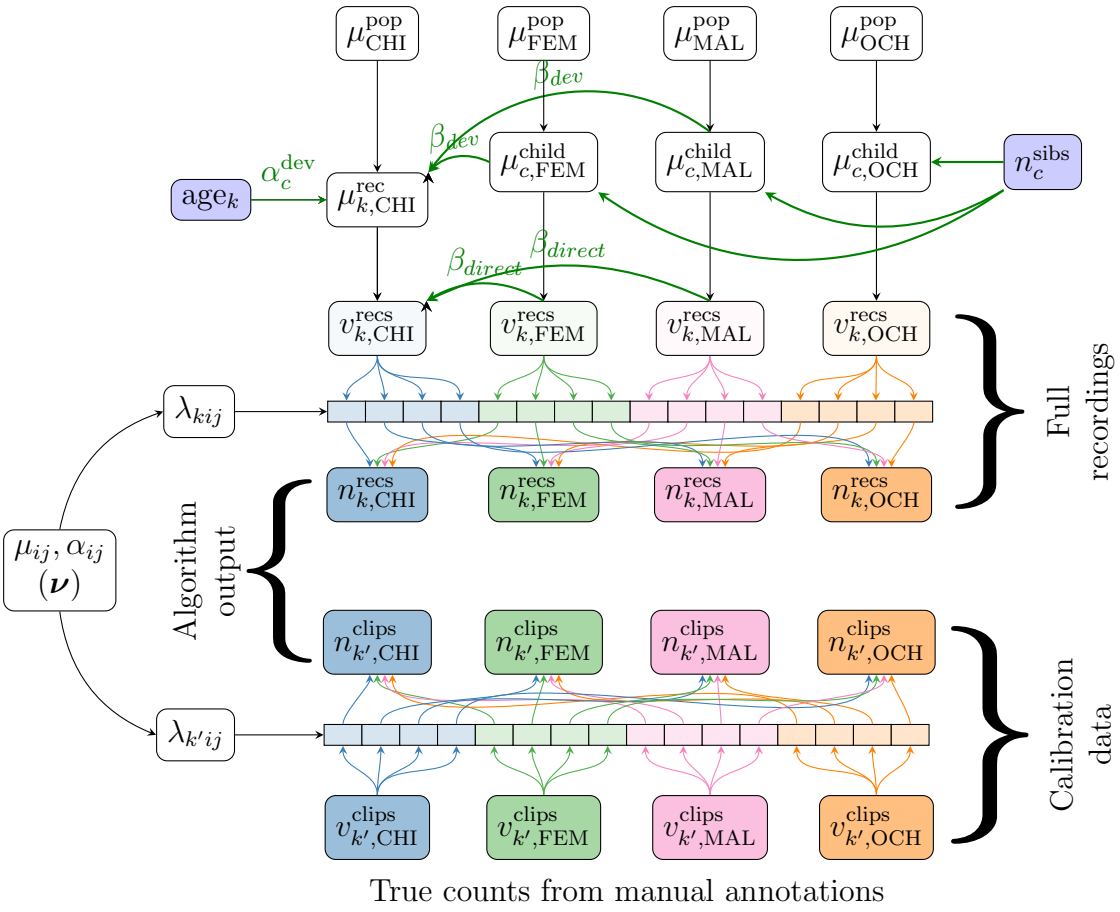


Figure 4: Combined model of speech behavior and of the algorithm behavior. Compared to Figure 3, the actual quantity of vocalizations (v^{recs}) is treated as latent variables. Colored arrows represent the effect of real vocalizations from each speaker (e.g. CHI, in blue) on the amount of vocalizations attributed to each speaker by the algorithm (n^{recs}). The unobserved confusion rates λ_{kij} represent the probability that vocalizations from a speaker i are detected and attributed to a speaker j in recording k. The distribution of λ_{kij}, parameterized by μ_{ij} and α_{ij}, is learned via calibration data (for which both the true counts n^{clips} and the algorithmic counts v^{clips} are known).

agreement between LENA™ and VTC on average (see Appendix A.2.3, Table 2). Most notably, the estimation of the amount of variance jointly explained (R^2) between the quantity of female adult speech measured by LENA™ and VTC increases from 0.62 to 0.73 after calibration. However, (α_{ij}) (the dispersion of the confusion rates) eventually remains difficulty to identify precisely, given the amount of manual annotations available, which was unfortunately bound by prior data collection efforts.

2.2.3 Simulations

Fitting the full model described above can be computationally challenging, and is sometimes unnecessary, given that certain effects are only marginally impacted by classification bias. In this context, simulations are a straightforward solution for assessing the sensi-

tivity of an analysis to classification errors (see Figure 5). To this end, the first step is to generate synthetic datasets reproducing the characteristics of the actual data to be analyzed (including the amount of participants and observations). Synthetic data is simulated by fixing the value of a parameter of interest, say $\hat{\theta}$ (for instance, by considering a null-hypothesis $\hat{\theta} = 0$). This gives $\hat{D}$, the synthetic "true" vocalization counts. The behavior of the algorithm itself is then simulated, yielding D_{meas} (the vocalization counts as they would be reported by the algorithm). The resulting synthetic datasets can then finally be processed within an analysis pipeline (e.g. a linear regression), producing an estimate of θ_{meas}. This estimate can be compared to the known true value ($\hat{\theta}$). This procedure can be repeated for different values of $\hat{\theta}$. If the difference between $\mathbb{E}(\theta_{\text{meas}})$ and $\hat{\theta}$ is generally negligible, then algorithmic errors do not seriously bias inferences and it is not necessary to perform any calibration. Whenever the difference is substantial enough to be problematic, we can employ the Bayesian calibration procedure described above (Section §2.2.2) for deriving unbiased estimates of θ. We apply this alternative approach to the measurement of the proportion of female adult speech in Section 3.4. Finally, we provide a Python package to facilitate this method (Gautheron, 2025).

$$\hat{\theta} \xrightarrow[\text{speech behavior}]{\text{simulate}} \hat{D} \xrightarrow[\text{algorithm}]{\text{simulate}} D_{\text{meas}} \xrightarrow[\text{analysis pipeline}]{\text{run}} \theta_{\text{meas}} \neq \hat{\theta}?$$

Figure 5: Process for assessing the impact of classification errors on a particular analysis pipeline. A true value $\hat{\theta}$ for a parameter of interest is drawn at random. The model of speech behavior is simulated given $\hat{\theta}$, which yields synthetic ground truth data, $\hat{D}$. The behavior of the algorithm is simulated, thus returning synthetic *algorithmic* data, $D \neq \hat{D}$. Finally, the analysis pipeline is run on D_{meas}. If θ_{meas} significantly departs from $\hat{\theta}$, then the process is sensitive to algorithmic bias.

2.3 Data

To inform the current exploration, we build on six corpora: Bergelson (English monolinguals, North America, Bergelson 2015), Cougar (English monolinguals, North America, VanDam 2018a), Lucid (English monolinguals, North America, Rowland et al. 2018), Kidd (English, predominantly monolinguals, Australia, Donnelly and Kidd 2021), Warlaumont (English-Spanish bilinguals, North America, Warlaumont et al. 2016), Winnipeg (mostly English monolinguals, with some French spoken, North America, McDivitt and Soderstrom 2016), Fausey (English monolinguals, North America, Mendoza and Fausey 2022; Fausey 2018). Dataset selection was constrained by a conjunction of imperatives: the data had to be longitudinal and to simultaneously contain the raw recordings, LENA and VTC annotations, and human annotations. Additionally, we only included audio data between 10am and 6pm, and we excluded recordings which do not cover this time

range entirely. This allows for more consistent comparisons across data points. The selected audio amounts to $\sim 11\,800$ hours total.

The two algorithms we consider differ on many aspects, from their training data to their architecture, but for the purposes of conciseness, we focus here on the aspect of their behavior most likely relevant to the question at hand: Their propensity to miss vocalizations, and confuse or co-activate speaker classes. LENA™ uses a Dirichlet Process Gaussian Mixture Model to classify audio frames into categories including key child speech, other speakers, noise, and silence. The system prioritizes precision over recall, meaning it tends to be conservative in identifying vocalizations to avoid false positives. VTC is an open-source neural network-based alternative to LENA that detects the same four speaker types but can handle overlapping speech and has been trained on multilingual data. Unlike LENA, VTC balances precision and recall by maximizing F-score, resulting in higher recall but lower precision. For detailed technical specifications and implementation details, see (Xu et al., 2008; Gilkerson et al., 2017) for LENA and (Lavechin et al., 2020, 2025) for VTC.

Corpus	Children	Recordings	Corpus age-range (mo)	Avg. child age-range (mo)	Human ann. (h)
bergelson	44	450	6-17	10.2	5.0
cougar	27	143	2-67	5.4	6.5
kidd	116	615	9-26	13.6	-
lucid	35	224	11-32	18.3	5.0
warlaumont	9	17	3-18	4.7	5.0
winnipeg	6	13	2-19	1.0	5.0
fausey-trio	0	0	-	-	1.1

Table 1: Corpora used in the present analyses. Fausey-trio is only used for purposes of calibration (full recordings from this corpus are not considered in the model).

Human annotation existed for snippets of five of our six key corpora, jointly covering only 0.23% of the total audio duration. Most of our human annotation data (totaling 20h) comes from the ACLEW collaboration, and has been documented in Soderstrom et al. (2021). Annotators were trained until they met stringent criteria on a gold standard, and all labs used the exact same annotation manual (the ACLEW DAS template, Casillas et al. 2017), including definitions of what constitutes a vocalization. We employed here the so-called "random sample": Fifteen two-minute clips were randomly sampled from one recording from each of 9-10 children in four English-spoken corpora. The use of random sampling safeguards against any bias in data selection coming from the use of an algorithm. An additional 6.5h come from the VanDam-5-minute corpus (VanDam, 2018b; Carns, 2015), which has been less overtly documented, but the description on Homebank (VanDam, 2018a) provides sufficient details by explaining that three non-consecutive

five-minute sections were sampled that had the highest child-adult conversational turns according to the automated LENA™ analysis. Annotators had access to LENA™ segmentation and could correct it, but did not have to, as their priority was to produce orthographic transcriptions of what was said. To further inform our analyses, we also included in-house human annotations on Fausey-Trio (Fausey, 2018), a seventh corpus that is otherwise not included in our analyses. Human annotation of this corpus was done independently of the present paper, with the purposes of contributing to a dataset with greater representation of male adults and other children. To this end, audio was sampled using 15-second long snippets and a loudness-based filter (i.e., independent from both LENA™ and VTC). Human annotators first listened to the snippets and only annotated them if there was at least one vocalization by adult males and/or other children (i.e., snippets with only key child and/or female adult vocalizations were not annotated). The annotation followed a simplification of ACLEW DAS focused only on segmentation, without transcription. In sum, most of the human annotation was done independently of the algorithms whose behavior is studied in this paper; and by design all four speaker classes are present in the human-annotated data. In total, the 27.6h of audio annotated by a human, VTC and LENA™ yielded 6638 × 15-second clips to be used for calibration purposes.[10]

3 Results

We now turn to our main goal, the study of the effect of classification errors on downstream analyses. First, in Section §3.1, we present evidence that classification errors distort correlations between the quantities of speech attributed to each speaker. Such distortions can bias our inferences about the relationship between input and output and distort our knowledge of several aspects of the input itself. In Section §3.2, building upon our calibration strategy and the calibration data, we compare the algorithms' confusion matrices. This will explain some of the discrepancies between the correlations estimated from the two algorithms under consideration. In Section §3.3, we compare the estimates of a variety of quantities and effect sizes relevant to language acquisition, using either manual annotations, or annotations from each algorithm, before and after applying our calibration strategy. Finally, in Section §3.4, we show how simulations can help anticipate and diagnose bias due to algorithmic errors.

[10]Note that to assess the impact of the recording environment on confusion rates (Appendix 20, Figure 20), we consider annotations from other corpora. To avoid reader confusion, those corpora and annotations are introduced in the relevant Appendix.

3.1 Correlations between speakers

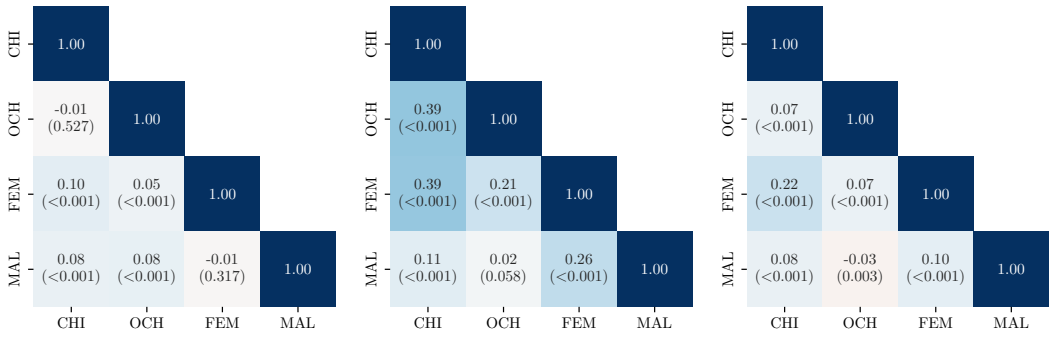
As shown in the illustration in Figure 2b, classification errors may open up biasing paths that produce spurious correlations between the quantity of speech attributed to different speakers. Using the calibration data (composed of 6638 × 15-second clips of audio annotated by a human, the VTC and the LENA), we test this hypothesis by comparing the correlation between each type of speaker across these clips according to each set of annotations (Figure 6). Manual annotations reveal statistically significant but low correlations between speakers across clips ($R \leq 0.10$). In contrast, VTC and LENA annotations produce much larger correlations, and are inconsistent with each other. For instance, the VTC reveals a large correlation between CHI and OCH ($R = 0.39$, $p < 0.001$), while the LENA finds a weak correlation ($R = 0.07$, $p < 0.001$), and manual annotations find no discernible correlation whatsoever ($R = -0.01$, $p = 0.527$). The LENA and the VTC also disagree strongly with each other and with manual annotations about the correlation between children's output and female adult input. This is problematic given the importance of the relationship between input and output in the context of language acquisition. All in all, Figure 6 clearly demonstrates the existence of biasing paths due to speaker misclassification previously illustrated in Figure 2b.

While revealing of classification errors, correlations between speech quantities measured in short clips bear little importance in themselves. Users are often more interested in correlations between vocalization counts aggregated at the level of whole recordings or the level of each child – those are shown in Figures 7 and 8 respectively. These correlations exhibit similar issues, with VTC and LENA™ reporting inconsistent correlations – particularly for those with vocalizations attributed to other children. This further suggests that the estimation of such correlations is affected by classification errors.

3.2 Confusion matrices of LENA™ and VTC

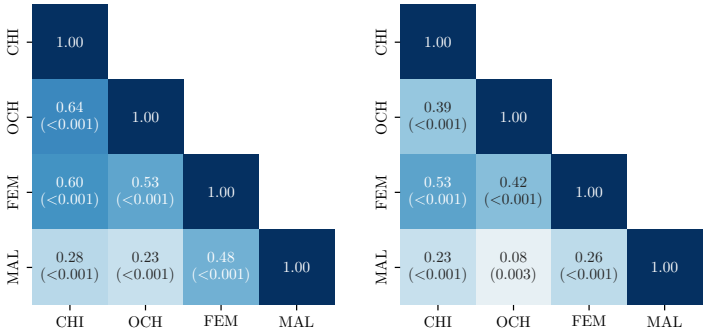
Below, we show that the rates of classification errors of LENA™ and VTC can explain the distortions of correlations between speakers shown in Figures 6–8.

Using the calibration model introduced in Section§2.2.2 together with the calibration data (i.e. clips of audio for which human annotations are available that provide a ground truth) reveals each algorithm's rates of true positive and false positive. These can be represented in the form of confusion matrices, as in Figure 9a. This shows that certain pairs of speakers are more often confused with one another: CHI and OCH (i.e. children) on the one hand, and FEM/MAL (i.e. adults) on the other hand. In addition, female adults are more often confused with children (and vice-versa) compared to male adults. Turning back to correlations in Figures 6–8, we indeed observe that VTC and LENA™ report higher correlations between CHI, OCH and FEM, MAL than humans. We also find that VTC and LENA™ report higher correlations between CHI/OCH and FEM than



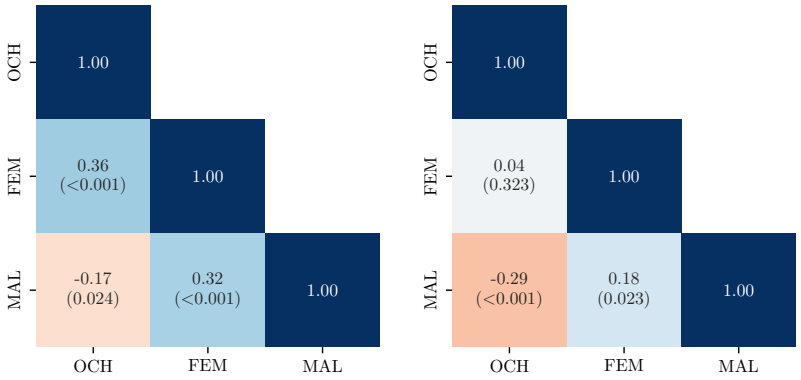
(a) Human annotations (b) VTC (c) LENA

Figure 6: Correlations in vocalization quantity across speakers, based on 6638×15s audio clips annotated by humans, the VTC, and the LENA.



(a) VTC (b) LENA

Figure 7: Correlation between the quantity of vocalizations attributed to each speaker across recordings. The correlation matrix is extracted using a hierarchical multivariate log-normal model described in Section §A.1.2. Estimates from manual annotations are not included due to a lack of data at the full-recording level.



(a) VTC (b) LENA

Figure 8: Correlation between the quantity of vocalizations attributed to each speaker across children. "CHI" is omitted since it varies significantly with age. The correlation matrix is extracted using a hierarchical multivariate log-normal model described in Section §A.1.2.

between CHI/OCH and MAL, again in line with what we expect from classification errors. In addition, the confusion matrix shows that VTC exhibits higher rates of false positives than LENA™ generally. Again, Figures 6–8 show that VTC reports higher correlations between speakers than LENA™ across the board, as expected from diarization errors.

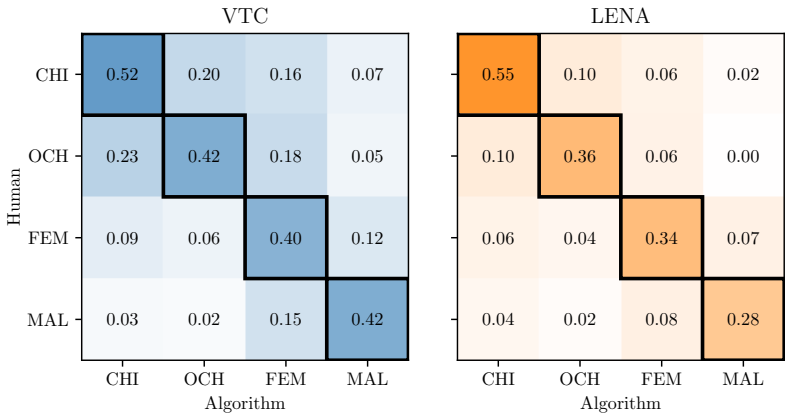
Although LENA™ has lower rates of false positives compared to VTC – in line with previous observations that LENA™ has a generally higher precision (Lavechin et al., 2025) (although it suffers from lower rates of true positives) consistent with the fact that VTC has a higher recall (Lavechin et al., 2025). Average confusion rates, however, are not sufficient metrics for comparing the merits of multiple classifiers. Another criterion is whether the performance of these algorithms is *consistent* throughout recording conditions. Imagine a classifier with a very stable detection rate of 50%. The actual quantity of events could still be estimated rather precisely, by multiplying the amount of detected events by two. A different classifier with a detection rate varying widely between 50 and 70% across recordings might be more "accurate" on average, but its calibration would be much trickier. To compare the stability of LENA™ and VTC, we show the distribution of confusion rates throughout recordings (Figure 9b). The distributions of the rates of true positives (on the diagonal) are more peaked for VTC, which indicate that the recall of LENA™ is considerably more variable than VTC.

Finally, shaded areas in Figure 9b shows the uncertainty about the underlying distribution of confusion rates, which is higher for under-represented speakers (especially male adults). This shows the importance of collecting and pooling as much calibration data as possible, possibly across multiple corpora. This calls for broader annotation efforts (Kunze et al., 2025) stimulated by standardized solutions for accumulating, storing, and sharing annotation data (Gautheron et al., 2022; ACLEW Consortium, 2024).

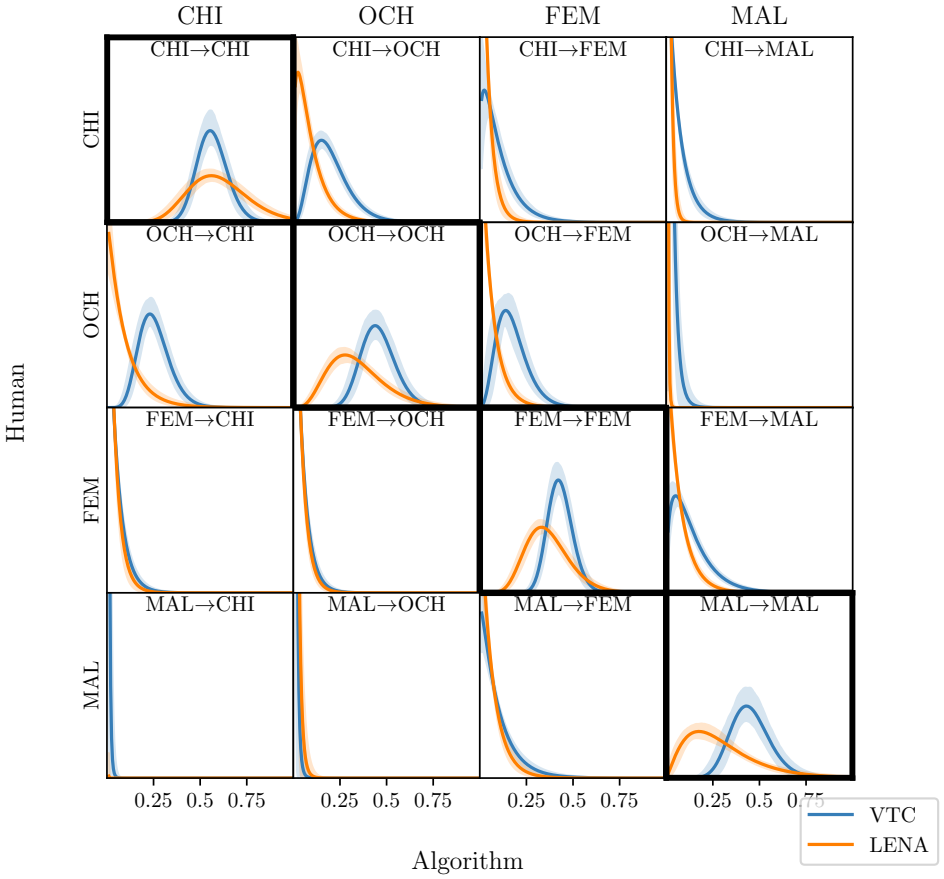
These findings are consistent with the distortions of correlations observed in the previous section. First of all, the VTC finds higher correlations across the board, which is consistent with its higher rates of false positives. In addition, regardless of the algorithm, the distortion of correlations is generally greater for pairs of speakers that are often confused with one another (e.g. CHI/OCH, OCH/FEM, or FEM/MAL). This is true for LENA™ as well, despite its lower rate of false positives. While this establishes the biasing effect of speaker confusion on correlations between speakers, the next section shows that this issue also affects measurements and inferences, implying quantities more obviously relevant to language acquisition.

3.3 Effect of classification bias on downstream analyses and measurements

In what follows, we report the effect of classification bias on six measurements of variables and effects relevant to language acquisition, using the model detailed in Section



(a) Average confusion rates (μ_{ij} in §2.2.2) of VTC and LENA™. Rows indicate the true speaker, and columns indicate the speaker class attributed by the algorithm. Diagonal elements represent the true positive rate for each speaker. Non-diagonal elements represent the distributions of the rates of false positives.



(b) Distribution of confusion rates across recordings, for VTC and LENA™ (i.e., the posterior distribution of λ_{kij}, cf. §2.2.2). Rows indicate the true speaker, and columns indicate the speaker class attributed by the algorithm. Diagonal elements represent the distributions of the rates of true positives for each speaker. Non-diagonal elements represent the distributions of the rates of false positives. Shaded areas represent the 95% credible interval of the error rates' probability density.

Figure 9: Confusion rates of VTC and LENA™.

§2.1 (see Figure 10). These are: (a) the proportion of female adult input; (b) the effect of age on children's speech output; (c) the effect of siblings on the quantity of input from other children and (d) from adults; (e) the direct effect of adult speech on children's output; (f) the long-term effect of adult input on children's output. Each measure is estimated using manual annotations alone, automated annotations without any calibration, and automated annotations with Bayesian calibration. Results are grouped by type of measurement.

In every case, the prior distribution (in black) reveals the initial hypothesis with respect to each measurement (the Bayesian prior), before considering any kind of data. The next set of observations (in gray) pertains to manual annotations, available only for portions of a limited subset of recordings. Based on such small quantities of data, these generally result in wide credible intervals, often largely overlapping with the prior (indicating that little to nothing could be learned). Automated annotations are associated with much narrower credible intervals in all cases, demonstrating the benefits of machine learning classifiers. However, additionally performing calibration on the automated annotations shifts these distributions in meaningful ways, which confirms the presence of misclassification bias across measurements. Post-calibration measurements are associated with larger credible intervals, reflecting our uncertainty in light of the stochastic nature of the algorithms' behavior.

What is the contribution of female adults to children's speech input? First, we consider the proportion of input attributed to female adult voices (10a). The prevalence of female versus male adult speech is relevant to a range of disciplines, feeding theoretical discussions on variation in parental investment as a function of family organization (e.g., Cassar et al., 2023) as well as interventions geared at greater involvement of fathers (e.g., Ferjan Ramírez et al., 2022). However, confusion between different speakers may distort our estimates of female input proportion (Figure 2a). For instance, vocalizations from children may be falsely attributed to adults; and vocalizations from male and female adults could be mistaken with each other, possibly at different rates. In Figure 10a, we find that LENA™ and VTC yield incompatible measurements of the proportion of female adult speech, with VTC probably underestimating it. After calibration, VTC and LENA™ estimates are closer to each other, and closer to the estimate achieved with manual annotations alone.

Do children vocalize more with age? Next, we consider the rate of increase in children's output with age (10b). Age-related increases in children's vocal production in long-form recordings have been widely documented, using both LENA™ (e.g. Bergelson et al., 2023) and VTC (e.g., Hervé et al., 2024). This is thought to reflect not simply maturational changes, but actually improvements in children's language skills, based on

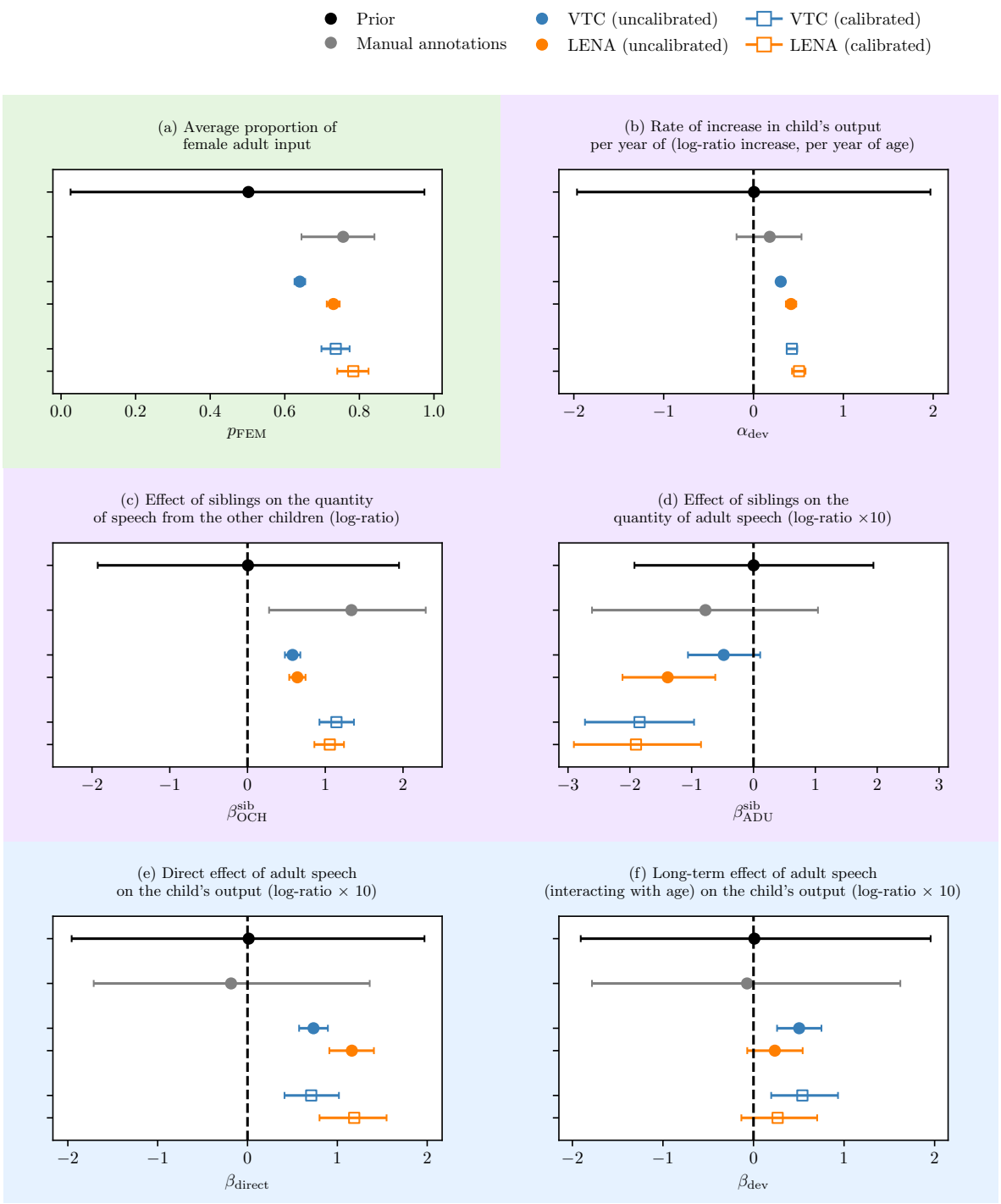


Figure 10: Comparison of effects' sizes derived with manual annotations alone (in gray) and automated annotations (in colors), without any calibration and with calibration. The prior distribution ($\mathcal{N}(0,1)$ or $\mathcal{U}(0,1)$, depending on the variable support) is shown in black for purposes of comparison. We distinguish three types of measurements: direct measurements of speech quantities (a); measurements of the effect of an independent variable on speech quantity (b, c, d); and measurements of the effect of a quantity of speech on another quantity of speech (e, f). Numerical values are reported in Appendix A.2.4, Table 3.

correlations between vocalization quantity and standardized measures of language development (a meta-analysis in Wang et al., 2020) and the observation that children with atypical development show a less pronounced age-related increase (Warlaumont et al., 2014b)[11]. Figure 10b shows that manual annotations are too sparse to confidently measure the increase in children's output with age. By contrast, automated annotations from LENA™ and VTC produce highly confident but non-overlapping estimates. Calibration increases the developmental effect of age, and reduces the gap between the two algorithms. Most likely, when classification errors are unaccounted for, speech output is contaminated with input speech that is less sensitive to the child's age, thus damping out the estimated effect of age.

How do siblings affect children's language input? Third, we consider the effect of siblings on the quantity of input children receive from other children (10c) and adults (10d). Among certain populations, speech from other children in the main exposure to language in infants (Cristia et al. 2023), and the contribution of siblings (as opposed to e.g. non-kins) may vary in rural and non-rural contexts. In our data, automated annotations without calibration show that children with no siblings receive about $\simeq 40\%$ fewer vocalizations from other children than children with siblings. This difference is implausibly low in the corpora under consideration in which siblings are expected to be, by far, the primary source of exposure to speech from other children. By contrast, the calibrated estimate ($\simeq 67\%$ less input from children for participants without siblings) is more plausible and more in line with the estimate derived from manual annotations alone. Without proper calibration, both algorithms under-estimate the effect size by a factor of two. Thus, without calibration, the contribution of non-kin children cannot be evaluated reliably, which is a problem for populations among which it is substantial. We then considered the effect of siblings on the input received from adults. Using one-hour long home-recorded videos, Laing and Bergelson (2024) find that children with more than one sibling receive less input from their caregivers. This result is consistent with "the resource dilution" model, a hypothesis that was put forward to explain a pattern of lower educational attainment as a function of sibship size by arguing that, as the number of children increases in a household, the main holders of intellectual resources (the adults) have to split them among the children, resulting in fewer resources per child (e.g., Kalmijn and van de Werfhorst, 2016). We tried to replicate this finding using our comprehensive automatic annotations of child-centered longform recordings. First of all, the effect of siblings on adult input is imperceptible with manual annotations alone. Automated annotations provide strikingly different estimates; while LENA™ finds that siblings reduce the amount of input afforded by adults to the child, VTC estimate is compatible with the null hypothesis (95% credible level). After calibration, the estimate

[11] Or even an age-related decrease (Hamrick et al., 2023).

of this effect is much larger, pointing to a 20% reduction in adult input among children with siblings, and consistent across VTC and LENA™.

Does hearing more speech make children talk more, now or in the long run? The final examples consider the measurement of the correlation between "input" (speech heard by children) and "output" (how much speech they produce) (10e, 10f). These have been previously approached in many different ways (as discussed in the meta-analysis by Anderson et al., 2021), including using daylong recordings (as in the meta-analysis by Coffey and Snedeker 2024). For example, Bergelson et al. (2023) reported a strong association between the quantity of adult speech afforded to children and the rate at which their speech production increases over time, a finding that is worth attempting to replicate. In the most diverse LENA™ study to date (Bergelson et al., 2023), input and output were drawn from the same audio-recordings, analyzed with LENA™ software, with CVC being predicted by adult vocalization counts (AVC). Of course, an association between input and output cannot be easily interpreted as indicating a causal relationship. For instance, (Bergelson et al., 2023) raise concerns a positive correlation may (partially) reflect shared genetics between the adults and children recorded[12]. As shown in Figure 2b, speaker classification errors could constitute an additional cause of confound, which, to our knowledge, has not been discussed. In our case, we distinguish the direct effect of input on output (surfacing as higher amounts of child speech in recordings with more input from adults, *ceteris paribus*), from the long-term developmental effect of input on child's speech, resulting from sustained exposure to higher input over time. No effects are found when using human annotations alone. Without calibration, VTC and LENA™ find positive effects, although their estimates are non-overlapping. After calibration, the effects remain roughly unaltered (except for larger uncertainties), and VTC/LENA™ continue to disagree. Persisting discrepancies between LENA™ and VTC indicate that the algorithms differ in ways unaccounted for by the calibration model. For instance, LENA™ does not handle overlapping speech, which makes the calibration model less suitable (due to the linear superposition assumption, cf. Section 2.2.2). In that respect, LENA™ is worse than VTC. The inability to support overlapping speech can distort correlations between speakers driven by very short-scale interactions. It is also more difficult to properly calibrate this model, because the errors of LENA™ are more dependent on the temporal distribution of speech.

In our typology of measurement, calibration works best for effects of a known variable on a speech quantity: intuitively, independent variables can help the model discriminate between spurious and actual correlations. Calibration was less useful for direct correlations between quantities of speech. One potential reason is that such correlations are

[12]This explanation would only be partial, since input is required to acquire language, and that input can have effects independent of shared genetics between mother and child (Huttenlocher et al., 2002).

most directly affected by classification errors and thus harder to tell apart, especially if the size of the actual effect is comparable in magnitude to the confusion rates. Finally, the approach may break-down for associations between speakers driven by interactions at very short-time scales and overlapping speech, for which the assumption of our model becomes invalid. This should also be a worry for researchers not using this approach, since it suggests that the downstream impact of algorithmic errors at such time-scales is difficult to evaluate.

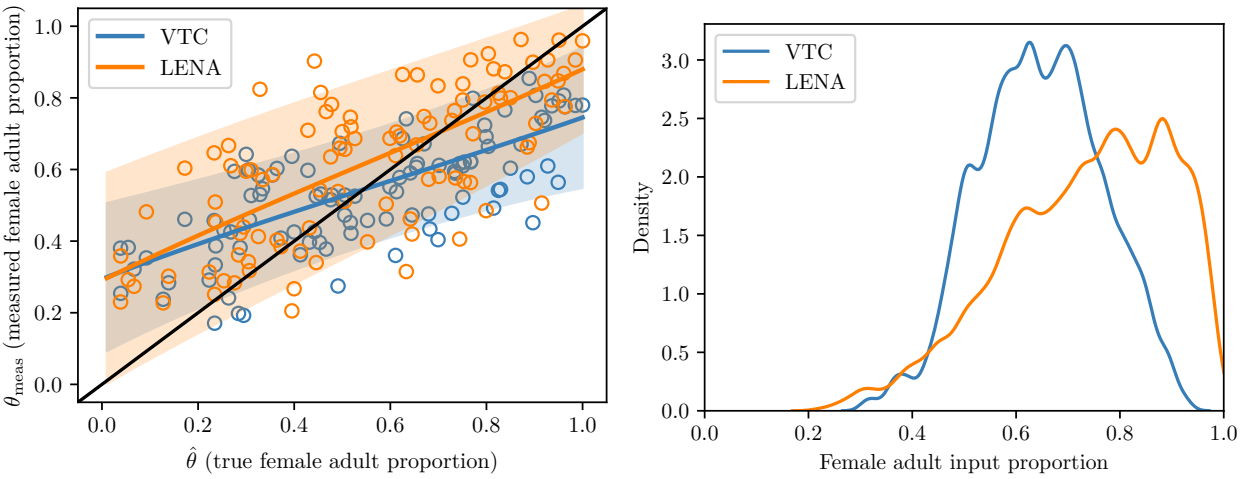
3.4 Anticipating biases with simulations

Classification bias can distort statistical measurements to varying extents. How to predict the level of sensitivity of a particular measurement to classification errors? This question can be answered with the simulation strategy proposed in Section §2.2.3. For instance, let us consider the relative contribution of female adults to adult input. We generate random datapoints representing the true amount of vocalizations from each speaker (CHI, OCH, FEM, and MAL) using the following generative process:

$$
\begin{aligned}
\text{CHI} &= 1500 \\
\text{OCH} &= \begin{cases} 0 & \text{with probability } 1/2 \\ 1000 & \text{with probability } 1/2 \end{cases} \\
\text{FEM} &= 3000 \times p \\
\text{MAL} &= 3000 \times (1-p) \\
p &\sim \text{uniform}(0,1)
\end{aligned}
$$

The total quantity of adult input (FEM+MAL) is fixed. The proportion of female adult input is represented by a parameter $\hat{\theta} = p$ drawn uniformly between 0 and 1. We can then simulate the algorithm's output for each generated sample, and compare the measured fraction of female adult speech θ_{meas} to the true value $\hat{\theta} = p$. The results are shown in Figure 11a (using 2000 samples). They demonstrate that estimates of the proportion of female adult speech are biased, especially for extreme proportions (close to 0 or 1). Moreover, LENA™ estimates higher fractions of female adult speech than VTC, and tends to overestimate female adult speech up to $p \sim 0.75$. This is probably because it was calibrated on training data dominated by female adult speech, together with the fact that the algorithm relies a lot on speakers' average prevalence to inform its classification boundary. Simulations also reveal the variance in measurements (as indicated by the shaded areas in Figure 11a). LENA™ has generally higher dispersion. The reliability of this simulation approach can be assessed by comparing these predictions to actual data.

Figure 11b shows the distribution of the proportion of female adult input estimated by VTC and LENA™ across the same corpus of audio recordings. The distribution obtained with LENA™ is shifted to the right (i.e., LENA™ reports higher proportions) and it has larger variance than the distribution given by VTC, as predicted by the simulations. Moreover, VTC saturates around $p = 0.9$, again in accordance with the simulations.



(a) True values versus measured values of female adult input proportion, according to simulations of each classifier (LENA and VTC). Markers show a fraction of the samples. Ideally, all points should fall onto the black line ($\theta_{\text{meas}} = \hat{\theta}$). Colored lines show the average trends. They depart from $\theta_{\text{meas}} = \hat{\theta}$, which implies the presence of bias. The shaded area represents the 90% probable interval.

(b) Distribution of the proportion of female adult input across recordings in the data, as estimated with VTC and LENA. LENA measures higher values than VTC, as predicted with simulations. The distributions, derived from the same recordings, are strikingly different.

Figure 11: Impact of speaker confusion on measurements of female and male adult speech.

4 Discussion

4.1 Summary

We investigated the statistical biases arising due to errors in automated classification. Although, we validated our approach on longform recordings in the context of language acquisition research, we note that it is conceptually applicable to any classification algorithm which detects events (in speech or other domains). We found significant discrepancies between the two dominant speech processing algorithm in our field (LENA™/VTC) on the very same audio ($\sim$ 1 400 recordings from 237 children across six longitudinal corpora). In themselves, these discrepancies across algorithms suggest that downstream measurements are distorted by classification errors. Additionally, we found evidence of spurious associations between speakers resulting from "biasing paths" (in a causal

inference language) opened up by speaker misidentification. Differences in correlation estimates across LENA™ and VTC seem consistent with what is expected given differences in these algorithms' confusion rates (e.g., the VTC has higher confusion rates, and reports higher correlations between speakers). Using simulations, we found even more evidence that classification errors can account for certain discrepancies between LENA™ and VTC. For instance, simulations informed by calibration data correctly predict that VTC underestimates the proportion of speech from female adults, more than LENA™, and produces less dispersed estimates across recordings.

To address these challenges, we developed a Bayesian calibration approach to remove biases resulting from the misidentification of speakers. This revealed that classification errors can lead to underestimating or overestimating measurements of quantities or effect sizes. Our Bayesian calibration strategy can identify and alleviate biases, while reducing the disagreement between different algorithms. It produces wider credible intervals, better reflecting our true uncertainty about our measurements. In the case of the effect of adult input on children's output, this approach fails to resolve the gap between VTC and LENA™. This suggests that there remain differences in these algorithms unaccounted for by the model. Most likely, these stem from differences in their behavior at short timescales, via distortions of the turn-taking patterns that may be driving the relationship between adult input and children's output. Given the importance of conversational patterns for language acquisition research, it is crucial to further investigate their distortions by diarization algorithms.

Below, we unpack the implications of our work for statistical inference with machine learning classifiers in behavioral science at large (Section §4.2). Finally, we discuss the implications for child-development/language acquisition research in particular (Section §4.3).

4.2 General implications for statistical inference with machine learning classifiers

4.2.1 Advantages and limitations of Bayesian calibration

Consistent with prior work, we find that more robust results are obtained through the *combination* of human and automated annotations (Angelopoulos et al., 2023b; TeBlunthuis et al., 2024). Indeed, most effects require too much data to be observed with human annotations alone, but automated annotations in isolation produce biased estimates. Bayesian calibration leverages both human and automated annotations to produce unbiased estimates with credible intervals reflecting our true uncertainty given the stochasticity of the underlying machine learning algorithms. In contrast to prediction-powered inference (Angelopoulos et al., 2023b), our Bayesian calibration approach is effective even with very few human annotations (0.2% of the audio in our case). In addi-

tion, our solution is flexible, since it can be developed independently from the behavioral model under investigation. Interestingly, Bayesian calibration can also flexibly integrate predictions from different classifiers into a single analysis, whether these cover disjoint or overlapping portions of the data[13]. However, in contrast to Angelopoulos et al. (2023b), our proposal requires assumptions about the behavior of the algorithm itself. Since these assumptions can be too simplistic or incorrect, this provides less guarantee in general.

We would like to suggest an area of improvement of our method that could yield benefits in a wide array of situations, far beyond this case study. As we have seen, LENA™ and VTC behave differently in part due to a different approach towards the precision/recall trade-off. The optimal balance of recall and precision might depend on the task at hand, and the relative harm of false positive versus false negatives (Silva Filho et al., 2023). We suggest that computer scientists allow the users of their algorithms to re-adjust the recall/precision trade-off according to their need, based on the confidence scores of the predictions[14]. More interestingly, confidence scores[15] could also be used as covariates in our calibration model, reducing the uncertainty in posterior statistical estimates. In the case of audio data, automatically inferred estimates of audio quality (e.g. signal-to-noise ratios or reverberation levels) are also correlated with accuracy and could act as complementary covariates in the calibration approach (Lavechin et al., 2023; Kunze et al., 2025).

4.2.2 Recommendations for statistical inference in behavioral studies

Although our main goal in this paper is to lay out some ways in which classifier errors may affect scientific conclusions rather than provide a fool-proof, general solution, we feel that it would be inappropriate to end the paper without attempting to produce a set of actionable recommendations.

First of all, standard measures in psychology such as accuracy, reliability, and validity are insufficient for anticipating and correcting classification bias. In contrast, causal graphical models are useful, both conceptually and practically: they enable us to track down the origin of potential biases, and they naturally imply statistical approaches to their resolution (Pearl, 2010). In addition, we recommend the recourse to simulations (leveraging causal graphical models such as ours) for assessing the impact of algorithmic errors on a measurement of interest, as we have illustrated with the proportion of female adult speech. Simulations are rather computationally inexpensive and technically easy

[13]In the latter case, one must however acknowledge potential correlations in the errors made by different algorithms.

[14]For instance, VTC computes an evidence score for each class (each speaker) at each audio frame, which is in a number between 0 and 1. In a second step, VTC decides if a class is active or not depending on whether the evidence score exceeds a certain threshold (which, in the case of VTC, is chosen to maximize the F-score, but could also be optimized differently)

[15]Whether or not these have been calibrated themselves, in the sense of Guo et al. 2017.

to implement. In the event that simulations reveal significant bias due to classification errors, our Bayesian calibration strategy can recover unbiased estimates, at the expense of higher computational costs. For the specific case of LENA™ and VTC, we provide a Python package simulating their measurements of vocalization counts from synthetic ground truth data (Gautheron, 2025).

As an alternative to simulations, it is possible to run an analysis pipeline using annotations from different classifiers. This can provide some indication about the sensitivity of the analysis to classification errors. For instance, LENA™ and VTC often yield non-overlapping estimates, given the differences in how they balance recall and precision (see Figure 9a). Additionally, they generally distort estimates in the same direction, which means that disagreement between LENA™ and VTC suggests that both are biased. This strategy, however, is not nearly as reliable as the appeal to simulations, because multiple classifiers can misbehave in similar ways, achieving mutually compatible but nevertheless biased estimates (see, e.g., Figure 10c).

4.3 Implications for child development and language acquisition research

4.3.1 Re-assessing LENA™ and VTC

Besides the observation that statistical measures of interest for child development are distorted by classification errors, this work offers a new opportunity to assess the relative merits of LENA™ and VTC. On average, due to a stronger emphasis on precision with respect to recall, LENA™ may seem less subject to classification bias. However, LENA™ performs notably worse for certain speaker types (OCH/MAL). In addition, its behavior is more variable, which increases measurement noise and renders calibration more challenging. Finally, certain aspects of LENA™, such as its inability to support overlapping speech, are harder to model and account for rigorously. This is an issue beyond the implementation of our approach; this also makes it more challenging to form expectations about the distortions LENA™ may introduce in statistical and causal inferences generally. For example, LENA™ is more likely to under-report child speech in the vicinity of adult speech. This has the potential to affect our estimate of the correlation between input and output. Eventually, in contrast to VTC, LENA™ fails to find evidence for a positive long-term effect of adult input on children's speech production in our data (at the 95% credible level). Finally, VTC is an open-source algorithm implemented in a flexible framework. It actively benefits from improvements incorporating state-of-the-art solutions in speech processing (Kunze et al., 2025). We hasten to indicate that two of the authors of the present paper are involved in both the original VTC and ongoing work, which may constitute a conflict of interest biasing our perception. We thus strongly encourage readers to peruse our results carefully to make up their own minds.

4.3.2 Issues for future work

The latent variable modeling approach to measurement calibration is promising, but several challenges must be addressed. First of all, it is crucial to better locate the sources of variation in the confusion rates: do they stem from variations within, or between recordings (or even systematic biases persisting throughout repeated observations of the same child)? Answering this question would require larger amounts of human annotations, covering i) more extensive fractions of the recordings and occasionally ii) multiple recordings from the same participants. Additionally, more overlap between annotators would be necessary to properly manage the effect of human errors, which we have neglected. Collecting these annotations at scale with the goal of calibration in mind may require more cooperation among researchers. In fact, the present work was only made possible by prior coordinated efforts of a network of scientists (ACLEW Consortium, 2024).

In addition, some researchers are studying properties of conversations (Abney et al., 2016), such as the probability of adults reinforcing children's speech-like, as opposed to non-speech, vocalizations (Warlaumont et al., 2014a); or even more fine-grained, studying the timing of inter-speaker turns (Ritwika et al., 2025). Extending our models above to say something about the temporal nature of speech in context will be mathematically challenging. In our Bayesian approach, this would require estimating the probability that a true sequence of vocalizations (e.g. adult, child, adult, child, ...) is detected by the algorithm as any other sequence (e.g., adult, adult, child, ...). The space of possible sequences is combinatorially large, and this may only work for short sequences. Time-coding precision, which we have ignored in this analysis, becomes important. Finally, algorithms distort sequences in complex ways, related to constraints such as the minimum duration of vocalizations in LENA™ and its inability to handle overlaps, or the fact that VTC overproduces vocalizations with durations that are multiples of 250ms. Characterizing the effect of such distortions is challenging, but nevertheless important for certain applications.

5 Conclusions

With this paper, we aimed to bring attention to the potential downstream consequences of classification and segmentation errors made by machine learning algorithms. As research attempts to capture behavior through denser and more ecological datasets, machine learning will become an unavoidable tool. We thus do not recommend abandoning it, but rather increasing our awareness of where and how this tool's imperfection can affect our scientific conclusions. This is but a first step in this direction, which we hope will quickly become outdated as others further our approach and explore alternative strategies.

Acknowledgements We are grateful to Richard McElreath for in-depth discussions and advice about our modeling strategy early-on in the project. We would like to thank Mark Vandam for providing additional data and information about the cougar corpus. We must also acknowledge Meg Cychosz, who suggested the idea examining the proportion of female adult speech. Finally, we are thankful to Riccardo Fusaroli, Camila Scaff, and Tarek Kunze for their feedback on this manuscript, and to DARCLE for the opportunity to present an early report of our findings.

Funding This work was granted access to the HPC resources of IDRIS under the allocation 2024-AD011012186 made by GENCI. This research was also undertaken with the assistance of resources from the National Computational Infrastructure (NCI Australia), an NCRIS enabled capability supported by the Australian Government. L.G. acknowledges funding from the DFG Research Training Group 2696. E.K. acknowledges funding from the Australian Research Council (CE140100041). M.L. acknowledges funding from The Simons Foundation International (034070-00033). A.C. acknowledges the J. S. McDonnell Foundation Understanding Human Cognition Scholar Award and European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ExELang, Grant agreement No. 101001095). This work would not have been possible without the ACLEW project, for which we acknowledge Agence Nationale de la Recherche (ANR-16-DATA-0004 ACLEW).

Conflicts of interests The authors declare no competing interests.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Availability of data and material The datasets analysed during the current study are not publicly available. However, the code includes synthetic data and routines to generate synthetic datasets, which allows readers to check the correctness of their implementation. In addition, the code can be run on any dataset formatted according to the ChildProject guidelines (Gautheron et al., 2022).

Code availability The code of our models and analyses is available at: https://gin.g-node.org/LAAC-LSCP/speaker-confusion-model [PENDING PERMANENT URL]. The Python simulation package is available at https://github.com/LAAC-LSCP/diarization-simulation.

Authors' contributions L.G.: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft; E.K.: Resources, Writing – review & editing; A.M.: Data curation, Writing – review & editing; M.L.: Writing – review & editing; A.C.: Writing – original draft, Supervision, Funding acquisition, Project administration.

References

Abney, D. H., Warlaumont, A. S., Oller, D. K., Wallot, S., & Kello, C. T. (2016). Multiple Coordination Patterns in Infant and Adult Vocalizations. *Infancy*, *22*(4), 514–539. https://doi.org/10.1111/infa.12165

ACLEW Consortium. (2024). ACLEW: Analyzing Child Language Experiences around the World [A Digging Into Data Project].

Anderson, N. J., Graham, S. A., Prime, H., Jenkins, J. M., & Madigan, S. (2021). Linking quality and quantity of parental linguistic input to child language skills: A meta-analysis. *Child Development*, *92*(2), 484–501.

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., & Zrnic, T. (2023a). *ppi-py: A python package for scientific discovery using machine learning* [Python package for prediction-powered inference]. https://github.com/aangelopoulos/ppi_py

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., & Zrnic, T. (2023b). Prediction-powered inference. *Science*, *382*(6671), 669–674. https://doi.org/10.1126/science.adi6000

Bastianello, T., Lorenzini, I., Nazzi, T., & Majorano, M. (2023). The Language ENvironment Analysis system (LENA): A validation study with Italian-learning children. *Journal of Child Language*, 1–21.

Bergelson, E. (2015). HomeBank English Bergelson Seedlings Corpus. https://doi.org/10.21415/T5PK6D

Bergelson, E., Soderstrom, M., Schwarz, I.-C., Rowland, C. F., Ramírez-Esparza, N., R. Hamrick, L., Marklund, E., Kalashnikova, M., Guez, A., Casillas, M., et al. (2023). Everyday language input and production in 1,001 children from six continents. *Proceedings of the National Academy of Sciences*, *120*(52), e2300671120.

Blackwell, M., Honaker, J., & King, G. (2017). A unified approach to measurement error and missing data: Details and extensions. *Sociological Methods & Research*, *46*(3), 342–369. https://doi.org/10.1177/0049124115589052

Bruyneel, E., Demurie, E., Boterberg, S., Warreyn, P., & Roeyers, H. (2021). Validation of the Language ENvironment Analysis (LENA) system for Dutch. *Journal of Child Language*, *48*(4), 765–791.

Carns, K. (2015). *Question exposure and production in preschoolers who are hard-of-hearing* [Masters thesis]. Washington State University master's thesis. http://vandammark.com/WSU/Carns_2015.pdf

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, *76*, 1–32.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models* (2nd). Chapman & Hall/CRC.

Casillas, M., Bunce, J., Soderstrom, M., Rosemberg, C., Alam, M. M. F., Stein, A., & Garrison, H. (2017). *Tutorials: Using the ACLEW DAS template*. https://osf.io/b2jep/wiki/home/

Cassar, A., Cristia, A., Grosjean, P., & Walker, S. (2025). It makes a village: child care and prosociality. *Journal of Economic Growth*. https://doi.org/10.1007/s10887-025-09254-6

Cassar, A., Cristia, A., Grosjean, P. A., & Walker, S. (2023). It makes a village: Allomaternal care and prosociality.

Coffey, J., & Snedeker, J. (2024, January). Does Talking To Children Matter? A Meta-Analysis. osf.io/aydcf

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.

Cristia, A., & Casillas, M. (2020). LENA recordings gathered from children growing up in Rossel Island.

Cristia, A., Gautheron, L., & Colleran, H. (2023). Vocal input and output among infants in a multilingual context: Evidence from long-form recordings in Vanuatu. *Developmental Science*, *26*(4). https://doi.org/10.1111/desc.13375

Cristia, A., Gautheron, L., Zhang, Z., Schuller, B., Scaff, C., Rowland, C., Räsänen, O., Peurey, L., Lavechin, M., Havard, W., et al. (2024). Establishing the reliability of metrics extracted from long-form recordings using LENA and the ACLEW pipeline. *Behavior Research Methods*, 1–20.

Cristia, A., Lavechin, M., Scaff, C., Soderstrom, M., Rowland, C., Räsänen, O., Bunce, J., & Bergelson, E. (2021). A thorough evaluation of the Language Environment Analysis (LENA) system. *Behavior research methods*, *53*, 467–486.

Donnelly, S., & Kidd, E. (2021). The longitudinal relationship between conversational turn-taking and vocabulary growth in early language development. *Child Development*, *92*(2), 609–625. https://doi.org/10.1111/cdev.13511

Efron, B. (1986). Double Exponential Families and Their Use in Generalized Linear Regression. *Journal of the American Statistical Association*, *81*(395), 709–721. https://doi.org/10.1080/01621459.1986.10478327

Fausey, C. (2018). HomeBank English Fausey Trio Corpus. https://doi.org/10.21415/T5JM4R

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*, *27*(2), 301–323.

Gautheron, L. (2025). *Diarization simulation: A python package for simulating speaker diarization with LENA and VTC from ground truth vocalization data*. https://github.com/LAAC-LSCP/diarization-simulation

Gautheron, L., Rochat, N., & Cristia, A. (2022). Managing, storing, and sharing long-form recordings and their annotations. *Language Resources and Evaluation*, *57*(1), 343–375. https://doi.org/10.1007/s10579-022-09579-3

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough Oller, D., Hansen, J. H., & Paul, T. D. (2017). Mapping the early language environment using all-day recordings and automated analysis. *American journal of speech-language pathology*, *26*(2), 248–265.

Gilkerson, J., Zhang, Y., Xu, D., Richards, J. A., Xu, X., Jiang, F., Harnsberger, J., & Topping, K. (2015). Evaluating language environment analysis system performance for Chinese: A pilot study in Shanghai. *Journal of Speech, Language, and Hearing Research*, *58*(2), 445–452.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *International conference on machine learning*, 1321–1330.

Hamrick, L. R., Seidl, A., & Kelleher, B. L. (2023). Semi-Automatic Assessment of Vocalization Quality for Children With and Without Angelman Syndrome. *American Journal on Intellectual and Developmental Disabilities*, *128*(6), 425–448.

Hervé, E., François, C., & Prevot, L. (2024). Daily auditory environments in French-speaking infants: A longitudinal dataset. *Workshop on Cognitive Modeling and Computational Linguistics*, 132–151.

Huttenlocher, J., Vasilyeva, M., Cymerman, E., & Levine, S. (2002). Language input and child syntax. *Cognitive Psychology*, *45*(3), 337–374. https://doi.org/10.1016/s0010-0285(02)00500-5

Kalmijn, M., & van de Werfhorst, H. G. (2016). Sibship size and gendered resource dilution in different societal contexts. *PloS one*, *11*(8), e0160953.

Kunze, T., Métais, M., Titeux, H., Elbert, L., Coffey, J., Dupoux, E., Cristia, A., & Lavechin, M. (2025). Challenges in Automated Processing of Speech from Child Wearables: The Case of Voice Type Classifier [In press]. *Proceedings of Interspeech*. https://arxiv.org/abs/2506.11074

Laing, C., & Bergelson, E. (2024). Analyzing the effect of sibling number on input and output in the first 18 months. *Infancy*, *29*(2), 175–195. https://doi.org/10.1111/infa.12578

Laudańska, Z., Caunt, A., Cristia, A., Warlaumont, A. S., Patsis, K., Tomalski, P., Warreyn, P., Abney, D. H., Borjon, J. I., Airaksinen, M., et al. (2025). From data to discovery: Technology propels speech-language research and theory-building in developmental science. *Neuroscience & Biobehavioral Reviews*, 106199.

Lavechin, M., Bousbib, R., Bredin, H., Dupoux, E., & Cristia, A. (2020). An open-source voice type classifier for child-centered daylong recordings. *arXiv preprint arXiv:2005.12656*. https://arxiv.org/abs/2005.12656

Lavechin, M., Hamrick, L. R., Kelleher, B., & Seidl, A. (2025). Performance and biases of the LENA® and ACLEW algorithms in analyzing language environments in Down, Fragile X, Angelman syndromes, and populations at elevated likelihood for autism.

Lavechin, M., Métais, M., Titeux, H., Boissonnet, A., Copet, J., Rivière, M., Bergelson, E., Cristia, A., Dupoux, E., & Bredin, H. (2023). Brouhaha: multi-task training for voice activity detection, speech-to-noise ratio, and c50 room acoustics estimation. *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 1–7.

Lehet, M., Arjmandi, M. K., Houston, D., & Dilley, L. (2021). Circumspection in using automated measures: Talker gender and addressee affect error rates for adult speech detection in the Language ENvironment Analysis (LENA) system. *Behavior research methods*, *53*, 113–138.

Long, B., Goodin, S., Kachergis, G., Marchman, V. A., Radwan, S. F., Sparks, R. Z., Xiang, V., Zhuang, C., Hsu, O., Newman, B., et al. (2024). The BabyView camera: designing a new head-mounted camera to capture children's early social and visual environments. *Behavior Research Methods*, *56*(4), 3523–3534.

McDivitt, K., & Soderstrom, M. (2016). *McDivitt HomeBank Corpus* (tech. rep.). 10.21415/T5KK6G

McDonald, M., Kwon, T., Kim, H., Lee, Y., & Ko, E.-S. (2021). Evaluating the Language ENvironment Analysis System for Korean. *Journal of Speech, Language, and Hearing Research*, *64*(3), 792–808.

Mendoza, J. K., & Fausey, C. M. (2022). Everyday parameters for episode-to-episode dynamics in the daily music of infancy. *Cognitive Science*, *46*(8), e13178.

Miao, J., & Lu, Q. (2024). Task-agnostic machine-learning-assisted inference. *arXiv preprint arXiv:2405.20039*.

Pearl, J. (2010). On measurement bias in causal inference. *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, 425–432. https://ftp.cs.ucla.edu/pub/stat_ser/r357.pdf

Peurey, L., Lavechin, M., Kunze, T., Khentout, M., Gautheron, L., Dupoux, E., & Cristia, A. (2025). Fifteen years of child-centered long-form recordings: Promises, resources, and remaining challenges to validity. *arXiv preprint arXiv:2506.11075*.

Pustejovsky, J. E. (2024). Implementing Efron's double Poisson distribution in Stan [Accessed on May 19, 2025]. https://jepusto.com/posts/double-poisson-in-Stan/

Ritwika, V., Schneider, S., Lopez, L. D., Mai, J., Gopinathan, A., Kello, C. T., & Warlaumont, A. S. (2025). Burstiness and interpersonal foraging between human infants and caregivers in the vocal domain. https://doi.org/10.48550/ARXIV.2505.01545

Rowland, C. F., Bidgood, A., Durrant, S., Peter, M., & Pine, J. M. (2018). The Language 0–5 Project Corpus. https://doi.org/10.17605/OSF.IO/KAU5F

Salerno, S., Miao, J., Afiaz, A., Hoffman, K., Neufeld, A., Lu, Q., McCormick, T. H., & Leek, J. T. (2025). ipd: an R package for conducting inference on predicted data (J. Wren, Ed.). *Bioinformatics*, *41*(2). https://doi.org/10.1093/bioinformatics/btaf055

Scaff, C., Casillas, M., Stieglitz, J., & Cristia, A. (2023). Characterization of children's verbal input in a forager-farmer population using long-form audio recordings and diverse input definitions. *Infancy*, *29*(2), 196–215. https://doi.org/10.1111/infa.12568

Silva Filho, T., Song, H., Perello-Nieto, M., Santos-Rodriguez, R., Kull, M., & Flach, P. (2023). Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Machine Learning*, *112*(9), 3211–3260. https://doi.org/10.1007/s10994-023-06336-7

Soderstrom, M., Casillas, M., Bergelson, E., Rosemberg, C., Alam, F., Warlaumont, A. S., & Bunce, J. (2021). Developing a Cross-Cultural Annotation System and Meta-Corpus for Studying Infants' Real World Language Experience (J. P. Röer, Ed.). *Collabra: Psychology*, *7*(1). https://doi.org/10.1525/collabra.23445

TeBlunthuis, N., Hase, V., & Chan, C.-H. (2024). Misclassification in automated content analysis causes bias in regression. Can we fix it? Yes we can! *Communication Methods and Measures*, 1–22.

Textor, J. (2015). Drawing and analyzing causal dags with dagitty. *arXiv preprint arXiv:1508.04633*.

Textor, J., & Liskiewicz, M. (2012). Adjustment criteria in causal diagrams: An algorithmic perspective. *arXiv preprint arXiv:1202.3764*.

VanDam, M. (2018a). HomeBank English Cougar Corpus. https://doi.org/10.21415/T5WT25

VanDam, M. (2018b). VanDam Public 5-minute HomeBank Corpus. https://doi.org/10.21415/T5388S

Wang, Y., Williams, R., Dilley, L., & Houston, D. M. (2020). A meta-analysis of the predictability of LENA™ automated measures for child language development. *Developmental Review*, *57*, 100921.

Warlaumont, A. S., Pretzer, G. M., Mendoza, S., & Walle, E. A. (2016). *Warlaumont HomeBank Corpus* (tech. rep.). doi:10.21415/T54S3C

Warlaumont, A. S., Richards, J. A., Gilkerson, J., & Oller, D. K. (2014a). A Social Feedback Loop for Speech Development and Its Reduction in Autism. *Psychological Science*, *25*(7), 1314–1324. https://doi.org/10.1177/0956797614531023

Warlaumont, A. S., Richards, J. A., Gilkerson, J., & Oller, D. K. (2014b). A social feedback loop for speech development and its reduction in autism. *Psychological science*, *25*(7), 1314–1324.

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.

Xu, D., Yapanel, U., Gray, S., Gilkerson, J., Richards, J., & Hansen, J. (2008). Signal processing for young child speech language development. *First Workshop on Child, Computer and Interaction*.

A Supplementary materials

A.1 Models of speech behavior

A.1.1 Main model

For each recording k of child c, the vocalization counts are modeled as:

$$v^{\text{recs}}_{k,\text{CHI}} \sim \text{Gamma}\left(\alpha^{\text{CHI}}_{\text{child}}, \frac{\alpha^{\text{CHI}}_{\text{child}}}{\mu^{\text{rec}}_{k,\text{CHI}}}\right) \tag{5}$$

$$v^{\text{recs}}_{k,s} \sim \text{Gamma}\left(\alpha^{s}_{\text{child}}, \frac{\alpha^{s}_{\text{child}}}{\mu^{\text{child}}_{c,s}}\right), \quad s \in \{\text{FEM}, \text{MAL}, \text{OCH}\} \tag{6}$$

The child's expected vocalization rate incorporates age effects and the influence of adult speech:

$$\mu^{\text{rec}}_{k,\text{CHI}} = \mu^{\text{pop}}_{\text{CHI}} \exp\left(\alpha^{\text{dev}}_{c} \cdot \frac{\text{age}_k}{12} + \beta^{\text{dev}} \cdot \frac{\text{age}_k}{12} \cdot \frac{\mu^{\text{child}}_{c,\text{ADU}} - \mu_{\text{ADU}}}{\sigma_{\text{ADU}}} + \beta^{\text{direct}} \cdot \frac{v^{\text{recs}}_{k,\text{ADU}} - \mu^{\text{child}}_{c,\text{ADU}}}{\sigma^{\text{child}}_{c,\text{ADU}}}\right) \tag{7}$$

At the child level, for children with known sibling status:

$$\mu^{\text{child}}_{c,\text{OCH}} \sim \text{Gamma}\left(\alpha^{\text{OCH}}_{\text{pop},d}, \frac{\alpha^{\text{OCH}}_{\text{pop},S_c}}{\mu^{\text{pop}}_{\text{OCH}} \exp(S_c \beta^{\text{OCH}})}\right) \tag{8}$$

$$\mu^{\text{child}}_{c,s} \sim \text{Gamma}\left(\alpha^{s}_{\text{pop},d}, \frac{\alpha^{s}_{\text{pop},S_c}}{\mu^{\text{pop}}_{s} \exp(S_c \beta^{\text{ADU}}/10)}\right), \quad s \in \{\text{FEM}, \text{MAL}\} \tag{9}$$

where $S_c = 1$ if child has siblings, $S_c = 0$ otherwise.

For children with unknown sibling status, the model uses a mixture, marginalizing over the cases $S_c \in \{0, 1\}$, with probability p^{sibs} and $1 - p^{\text{sibs}}$ respectively.

The population-level parameters have the following priors:

$$\mu_s^{\text{pop}} \sim \text{Gamma}(2, 8) \quad \text{(prior mean = 250 vocs/hour)} \tag{10}$$

$$\alpha_{\text{pop},d}^{s} \sim \text{Gamma}(8, 1) \tag{11}$$

$$\alpha_{\text{child}}^{s} \sim \text{Gamma}(4, 1) \tag{12}$$

Developmental effects are modeled with:

$$\alpha_c^{\text{dev}} \sim \text{Normal}(\alpha^{\text{dev}}, \sigma^{\text{dev}}) \tag{13}$$

$$\alpha^{\text{dev}} \sim \text{Normal}(0, 1) \tag{14}$$

$$\sigma^{\text{dev}} \sim \text{Exponential}(1) \tag{15}$$

$$\beta^{\text{dev}} \sim \text{Normal}(0, 1) \tag{16}$$

$$\beta^{\text{direct}} \sim \text{Normal}(0, 1) \tag{17}$$

Sibling effects are captured by:

$$S_c \sim \text{Bernoulli}(p^{\text{sibs}}) \tag{18}$$

$$p^{\text{sibs}} \sim \text{Uniform}(0, 1) \tag{19}$$

$$\beta^{\text{OCH}} \sim \text{Normal}(0, 1) \tag{20}$$

$$\beta^{\text{ADU}} \sim \text{Normal}(0, 1) \tag{21}$$

Notation:

- $v_{k,s}^{\text{recs}}$: vocalization count for speaker s in recording k
- $\mu_{c,s}^{\text{child}}$: expected vocalization rate for speaker s for child c
- $\alpha_{\text{child}}^{s}$: variance parameter for speaker s at child level
- $\alpha_{\text{pop},d}^{s}$: variance parameter for speaker s at population level
- μ_s^{pop}: population-level average for speaker s
- α_c^{dev}: child-specific age effect
- β^{dev}, β^{direct}: developmental coefficients
- S_c: indicator for whether child c has siblings

A.1.2 Surrogate models for correlation estimates

In order to evaluate recording-level and child-level correlations between speakers, we use surrogate models as alternatives to the above model.

Recording-level correlations between speakers In this model for each recording k of child c, the vocalization counts are now modeled using a multivariate log-normal distribution:

$$\log \mathbf{v}_k^{\text{recs}} \sim \text{MVN}\left(\boldsymbol{\mu}_k - \frac{\text{diag}(\Sigma)}{2}, \Sigma\right) \tag{22}$$

where $\mathbf{v}_k^{\text{recs}} = (v_{k,\text{CHI}}^{\text{recs}}, v_{k,\text{OCH}}^{\text{recs}}, v_{k,\text{FEM}}^{\text{recs}}, v_{k,\text{MAL}}^{\text{recs}})$ is the vector of vocalization counts for all speaker types.

The mean vector $\boldsymbol{\mu}_k$ is defined as:

$$\mu_{k,\text{CHI}} = \log(\mu_{\text{CHI}}^{\text{pop}}) + \chi_k \tag{23}$$

$$\mu_{k,s} = \log(\mu_{c,s}^{\text{child}}), \quad s \in \{\text{OCH}, \text{FEM}, \text{MAL}\} \tag{24}$$

The child's developmental effect χ_k is:

$$\chi_k = \alpha_c^{\text{dev}} \cdot \frac{\text{age}_k}{12} + \beta^{\text{dev}} \cdot \frac{\text{age}_k}{12} \cdot \frac{\mu_{c,\text{ADU}}^{\text{child}} - \mu_{\text{ADU}}}{\sigma_{\text{ADU}}} \tag{25}$$

Note that this formulation removes the direct effect term (β^{direct}) that was present in the original model and introduces a covariance structure Σ between the different speaker types. The term $-\text{diag}(\Sigma)/2$ ensures the expected values after exponentiation match the desired means.

The covariance matrix Σ is parameterized through its Cholesky decomposition L_Σ such that $\Sigma = L_\Sigma L_\Sigma^T$.

All other aspects of the model (child-level parameters, population-level parameters, developmental effects, and sibling effects) remain unchanged from the original specification.

Child-level correlations between speakers In a third model aimed at measuring correlations between speakers at the child-level, the child-level parameters are assumed to follow a multivariate log-normal distribution:

$$\log(\boldsymbol{\mu}_c^{\text{child}}) \sim \text{MVN}\left(\log(\boldsymbol{\mu}_{2:n}^{\text{pop}}) - \frac{\text{diag}(\Sigma_{\text{child}})}{2}, \Sigma_{\text{child}}\right) \tag{26}$$

where $\boldsymbol{\mu}_c^{\text{child}} = (\mu_{c,\text{OCH}}^{\text{child}}, \mu_{c,\text{FEM}}^{\text{child}}, \mu_{c,\text{MAL}}^{\text{child}})$ and $\boldsymbol{\mu}_{2:n}^{\text{pop}} = (\mu_{\text{OCH}}^{\text{pop}}, \mu_{\text{FEM}}^{\text{pop}}, \mu_{\text{MAL}}^{\text{pop}})$.

The parameters for defining the adult speech influence are now:

$$\mu_{\text{ADU}} = \mu^{\text{pop}}_{\text{FEM}} + \mu^{\text{pop}}_{\text{MAL}} \tag{27}$$

$$\sigma_{\text{ADU}} = \sqrt{(\exp(\Sigma^{\text{child}}_{\text{FEM,FEM}}) - 1)(\mu^{\text{pop}}_{\text{FEM}})^2 + (\exp(\Sigma^{\text{child}}_{\text{MAL,MAL}}) - 1)(\mu^{\text{pop}}_{\text{MAL}})^2} \tag{28}$$

All other model components (developmental effects, population-level parameters) remain unchanged from the first model.

The key difference in this model is that it captures correlations between different speaker types at the child level through the multivariate log-normal distribution, while maintaining the original Gamma distributions for the recording-level observations.

A.1.3 Fitting the model on human annotations alone

Recordings for which human annotations are available are only partially annotated (typically 30 minutes of audio is annotated, out of many hours). We therefore make the assumption that the relationship between the manual vocalization counts ($\boldsymbol{n}^{\text{human}}_{\boldsymbol{k}} = (n^{\text{human}}_{k,CHI}, n^{\text{human}}_{k,OCH}, n^{\text{human}}_{k,FEM}, n^{\text{human}}_{k,MAL})$) and the unobserved vocalization counts for the whole recording ($\boldsymbol{v_k}$) is:

$$\boldsymbol{n}^{\text{human}}_{\boldsymbol{k}} \sim \text{Poisson}\left(\frac{\tau_{\text{annotated}}}{\tau_{\text{rec}}} \cdot \boldsymbol{v_k}\right) \tag{29}$$

Where $\tau_{\text{annotated}}$ is the duration of the audio that was hand-annotated. This makes the drastic (and false) assumption that the vocalization rate is constant throughout the recordings, leading to overconfident credible intervals. Thus, in reality, human annotations alone are even *less* informative then we report in Figure 10. The benefit of complementing human annotations with automated annotations is thus even larger.

A.2 Model of classification errors

A.2.1 Model validation

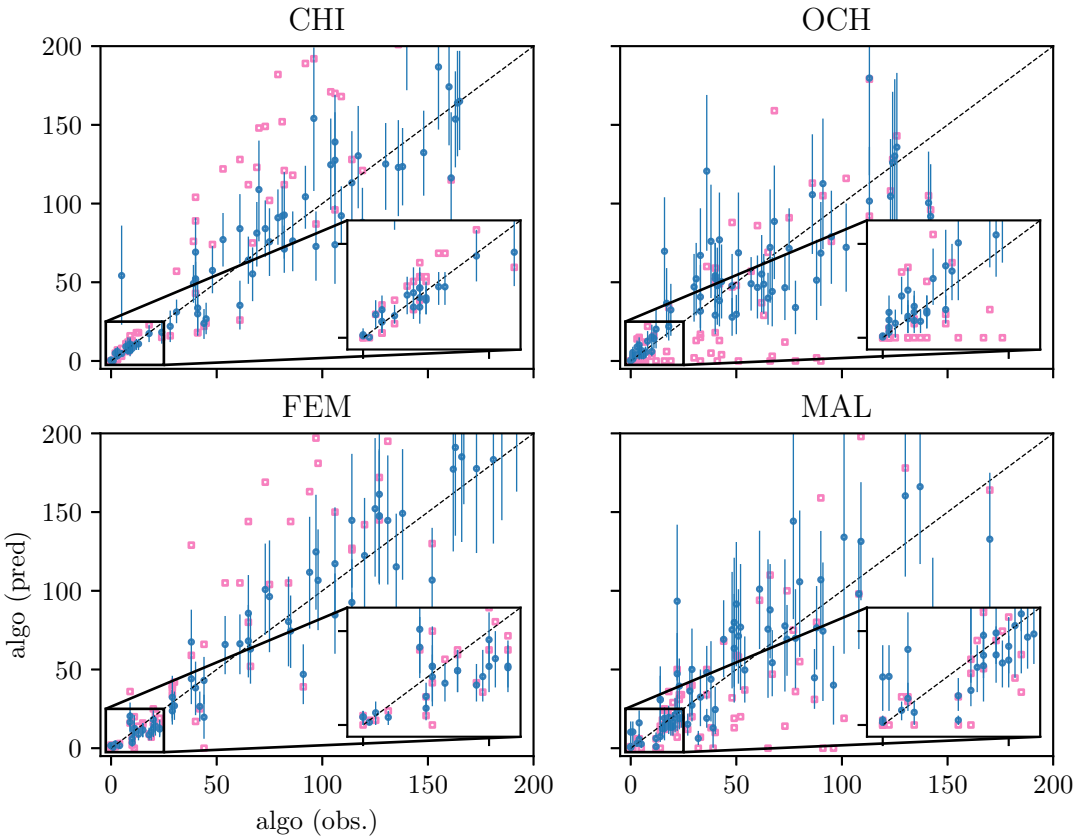


Figure 12: Relationship between the vocalization counts actually derived with VTC and the quantities expected to be derived from VTC given the model of the algorithm behavior and the true vocalization counts. Each blue point represents one of the recordings from the calibration data. The x-axis indicates the amount of vocalizations detected by the VTC for each speaker. The y-axis represents the amount expected based on the algorithm behavior (in blue) and the true amount of vocalizations for each speaker (in pink). The error bars indicate 68% probable intervals; most of the uncertainty lies in the variance in the confusion rates across recordings.

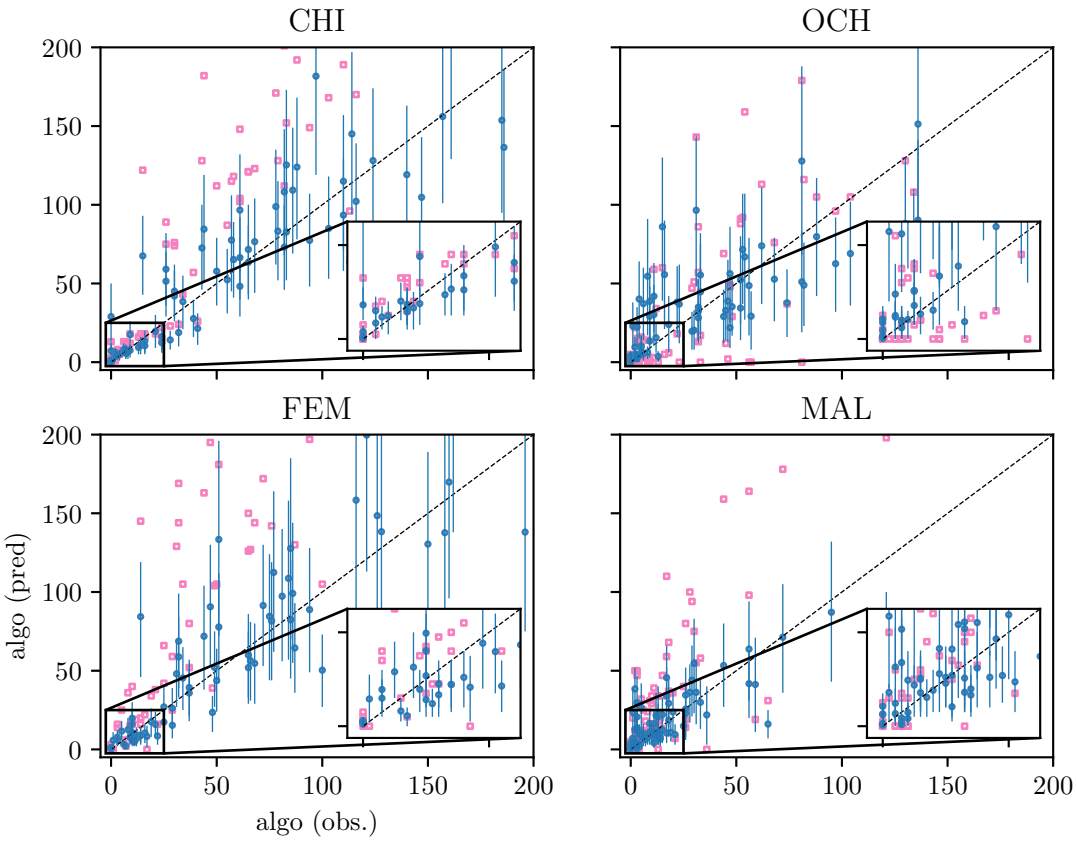


Figure 13: Relationship between the vocalization counts actually derived with LENA™ and the quantities expected to be derived from LENA™ given the model of the algorithm behavior and the true vocalization counts. Each blue point represents one of the recordings from the calibration data. The x-axis indicates the amount of vocalizations detected by LENA™ for each speaker. The y-axis represents the amount expected based on the algorithm behavior (in blue) and the true amount of vocalizations for each speaker (in pink). The error bars indicate 68% probable intervals; most of the uncertainty lies in the variance in the confusion rates across recordings.

A.2.2 Validation via simulations

To further validate our approach, we apply the model to simulated human and algorithmic annotations comparable in size to our calibration dataset. We draw vocalizations by assuming a Poisson process. Vocalization durations are drawn uniformly between 1s and 2s. Confusion rates across simulated recordings are drawn from a Beta distribution with mean μ_{ij} and shrinkage parameter $\eta = 50$. Simulations show that our approach is able to identify the correct confusion rates as long as speech density remains not too high (Figure 14).

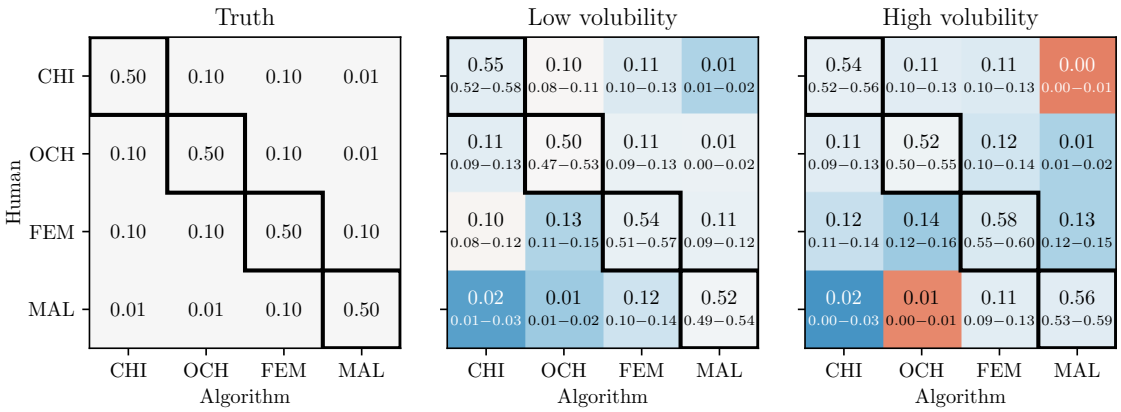


Figure 14: True confusion (μ_{ij}) compared to the confusion matrices recovered by the model from simulated data, under normal and high volubility. Colors indicate deviations from the true values (blue indicates overestimates, and red indicates underestimates). High volubility can lead to overestimating (μ_{ij}).

A.2.3 VTC/LENA comparison prior and after calibration

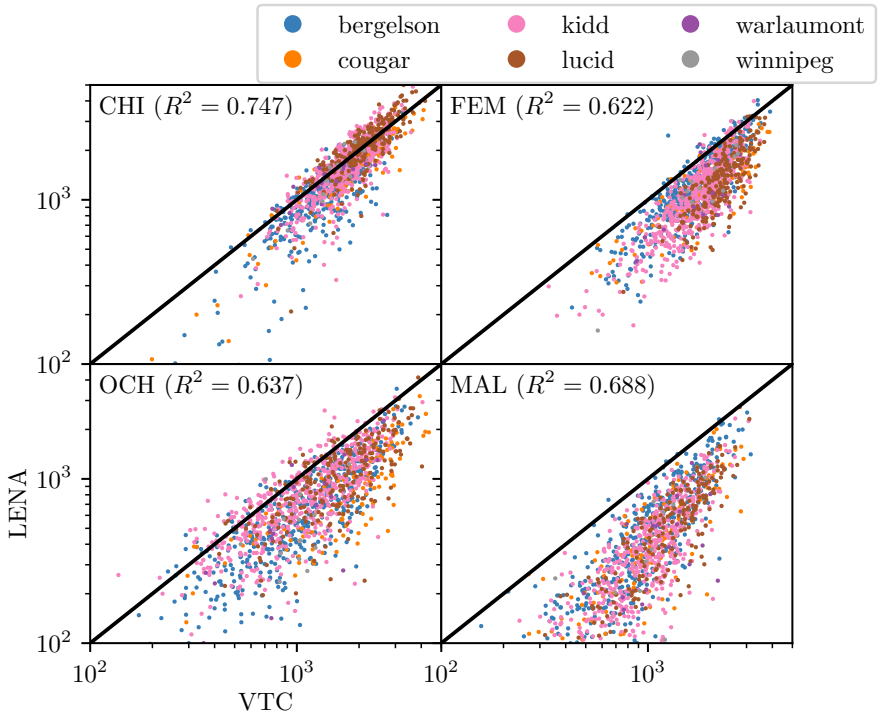


Figure 15: Comparison of vocalization counts derived with VTC (x-axis) and LENA™ (y-axis) per speaker and per recording, prior to any calibration.

	CHI	OCH	FEM	MAL
Before calibration	0.747	0.637	0.622	0.688
After calibration	0.744	0.713	0.733	0.744

Table 2: Correlation (R^2) between LENA and VTC measurements for each speaker, with and without calibration.

A.2.4 Results

		Manual annotations	Automated annotations			
			Prior to calibration		After calibration	
			VTC	LENA	VTC	LENA
Speech quantity	Female adult proportion	0.76 [0.64, 0.84]	0.64 [0.63, 0.65]	0.73 [0.71, 0.75]	0.74 [0.70, 0.77]	0.78 [0.74, 0.82]
Effect of independent variables on speech quantities	Age → output	0.18 [−0.19, 0.53]	**0.30** [**0.26**, **0.35**]	**0.42** [**0.36**, **0.48**]	**0.43** [**0.37**, **0.49**]	**0.51** [**0.43**, **0.58**]
	Siblings → input from children	**1.34** [**0.28**, **2.29**]	**0.58** [**0.48**, **0.68**]	**0.64** [**0.54**, **0.75**]	**1.14** [**0.93**, **1.37**]	**1.06** [**0.86**, **1.24**]
	Siblings → adult input	−0.78 [−2.61, 1.04]	−0.48 [−1.06, 0.11]	**−1.39** [**−2.12**, **−0.62**]	**−1.84** [**−2.73**, **−0.96**]	**−1.90** [**−2.91**, **−0.85**]
Associations between speech quantities	Input → output (direct)	−0.18 [−1.71, 1.36]	**0.73** [**0.57**, **0.89**]	**1.16** [**0.91**, **1.41**]	**0.71** [**0.41**, **1.02**]	**1.19** [**0.80**, **1.55**]
	Input → output (long-term)	−0.07 [−1.78, 1.62]	**0.50** [**0.26**, **0.75**]	0.24 [−0.07, 0.54]	**0.54** [**0.20**, **0.93**]	0.27 [−0.13, 0.70]

Table 3: Comparison of regression estimates from manual and automated annotations, including their posterior 95% credible intervals. Significant effects appear in bold characters.

A.2.5 Alternative approach (direct comparison)

An undesirable feature of the approach based on vocalization counts in 15s clips is that reliance on an arbitrary window-size. Shorter windows introduce boundary effects; longer windows decrease the informativeness of each individual clip, challenging the ability of the model to learn the distribution of confusion rates.

We considered an alternative strategy, based on an direct comparison of the vocalizations retrieved by the algorithms and the human annotators. In this strategy, an algorithmic vocalization attributed to a speaker i is considered a true positive iif it has a non-zero intersection with a real vocalization from the same speaker. It is considered a false positive iif it has no intersection with actual vocalizations from the correct speaker (i), but intersects with vocalizations from a single other speaker $j \neq i$. For each recording k with human annotations, we thus directly derive an estimate of n_{kij}, the amount of vocalizations attributed to j as a result of vocalizations from i. This approach, however, necessarily underestimates confusion rates; in particular, it fails to capture vocalizations misidentified every time a detected vocalization intersects with actual vocalization from two speakers. Figure 16 confirms that this yields lower confusion rates than the 15s clips method.

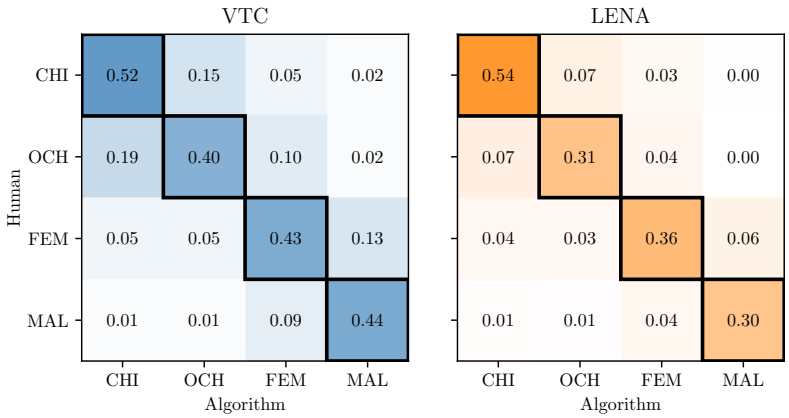


Figure 16: Average confusion rates of VTC and LENA™, through a direct comparison of human and algorithmic annotations. Rows indicate the true speaker, and columns indicate the speaker class attributed by the algorithm. Diagonal elements represent the true positive rate for each speaker. Non-diagonal elements represent the distributions of the rates of false positives.

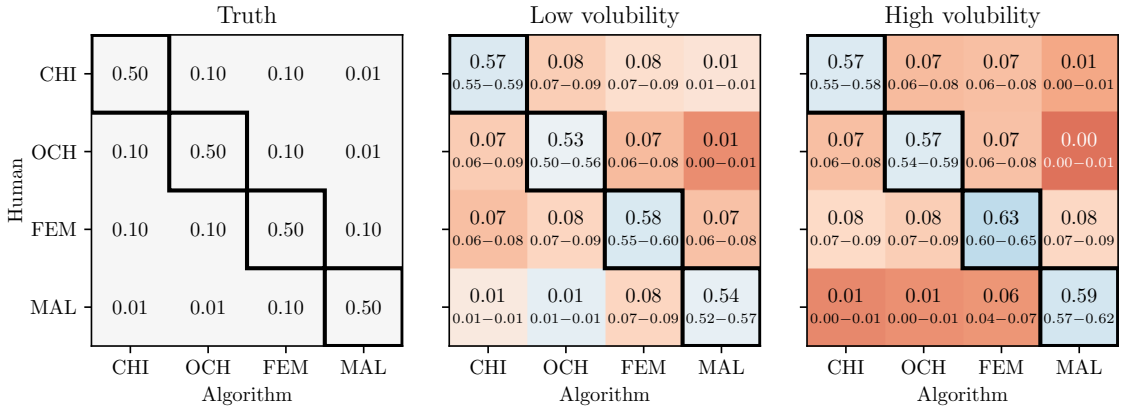


Figure 17: True confusion (μ_{ij}) compared to the confusion matrices recovered by the alternative inference strategy from simulated data, under normal and high volubility. The matrix to the left represents the true values. Colors indicate deviations from the true values (blue indicates overestimates, and red indicates underestimates). The alternative strategy generally underestimates misclassification errors and overestimates true positives.

A.2.6 Downstream comparison of the two calibration strategies

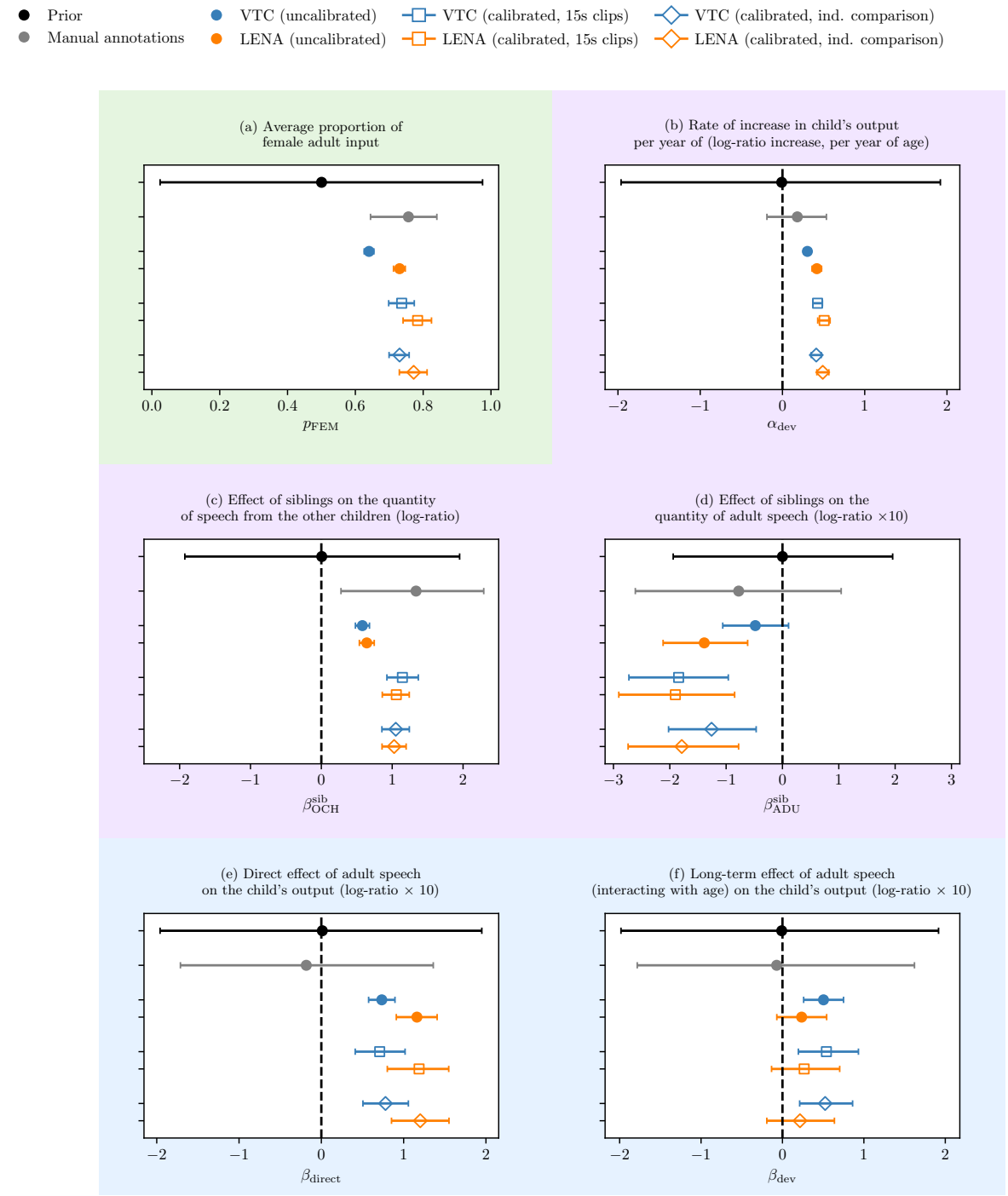


Figure 18: Comparison of effects' sizes derived with manual annotations alone (in gray) and automated annotations (in colors), without any calibration and with calibration. The prior distribution ($\mathcal{N}(0,1)$ or $\mathcal{U}(0,1)$, depending on the variable support) is shown in black for purposes of comparison. We distinguish three types of measurements: direct measurements of speech quantities (a); measurements of the effect of an independent variable on speech quantity (b, c, d); and measurements of the effect of a quantity of speech on another quantity of speech (e, f).

A.3 Effect of the child's age and environment on confusion rates

Figures 9a and 9b collapse across all children and corpora. In reality, the confusion matrix might depend on a number of factors, in which case collapsing across them is inappropriate. One of them is the child's age, which could affect the ability of the algorithm to correctly detect and classify children's vocalization. If that were the case, this could potentially undermine the ability of inferring the effect of age on the child's speech production. The effect of age on the algorithm performance on the vocalizations of the key child is shown for different age groups in Figure 19. There is some evidence that the detection rate for children increases after two years of age.

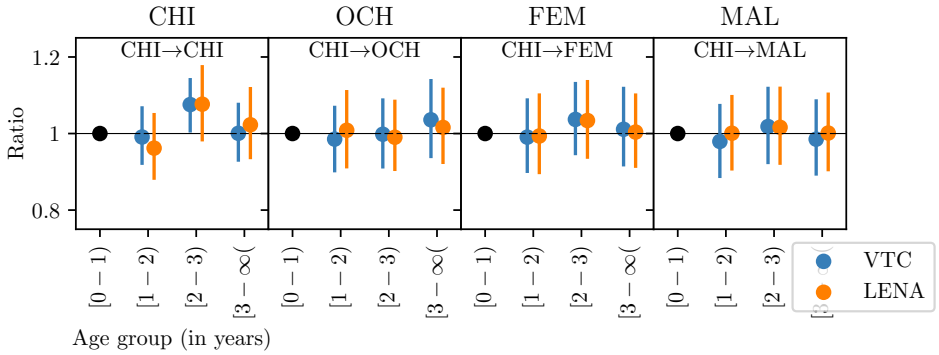


Figure 19: Effect of child age on the confusion rates $\lambda_{\text{CHI},i}$ (CHI→ $i \in$ {CHI,OCH,FEM,MAL}). The mean confusion rate for each age bin is compared the mean confusion rates for children between zero and one year of age (a ratio of one implies no difference). For VTC, variations larger than ∼10% in the mean true positive rates are excluded.

Besides child age, it is also conceivable that confusion rates depend on environmental factors (e.g. time spend outside, exposure to noise, etc.) and vary across languages. We therefore sought to compare the confusion rates for corpora drawing from urban English-speaking populations with confusion rates estimated from recordings of rural and non-English speaking populations. To this end, we drew from corpora that sampled "rural" populations in Papuasia New Guinea (Cristia & Casillas, 2020), the Solomon Islands (Cassar et al., 2025), Vanuatu (Cristia et al., 2023), and Bolivia (Scaff et al., 2023). Some of these recordings were done with devices other than LENA™. Moreover, all of these languages are considered under-resourced, which may mean that algorithms built on cumulative knowledge in the speech technology literature may be particularly ill-suited to them. In any case, English was seldom if ever spoken in these recordings, which makes them mismatch in training set with LENA™ (which was trained on North American English audio). As a result, our comparison conflates across many dimensions, all of which predict poorer algorithm performance in the "rural" recordings. Thus, while this comparison does not allow to isolate the effect of the environment – urban vs rural –, language, or recording device (which we cannot do due to limited sample size), it gives

an idea about whether any of these factors could alter classification errors.

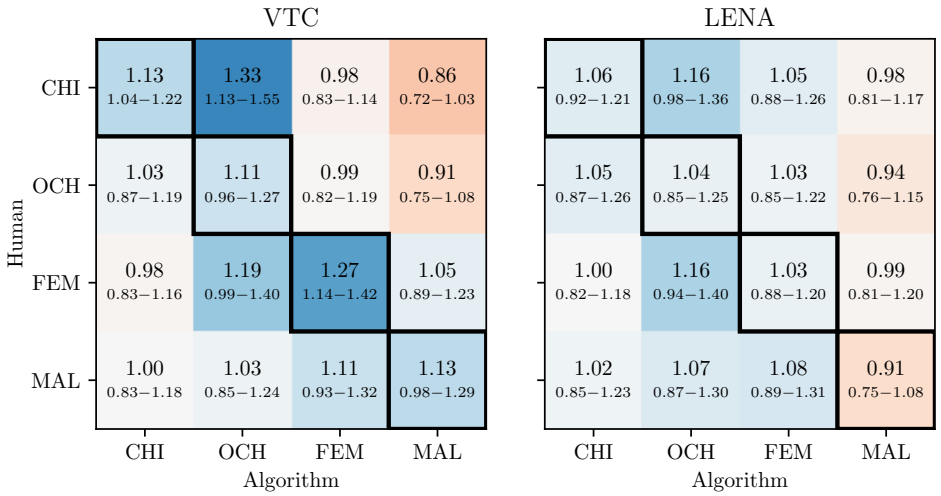


Figure 20: Rural/urban confusion rates for VTC (left) and LENA™ (right). The latter include non-English speaking, non-WEIRD populations and recorders other than LENA™. Values greater than one (blue cells) signal higher confusion rates among rural corpora than among urban corpora. 95% credible intervals are indicated underneath each value.

Figure 20 finds modest differences, most of them being statistically non-significant.

A.3.1 Stan parameters

	Chains	**Warmup iter.**	**Sampling iter.**	**Accept. delta**	**Max tree-depth**	**CPU cores**	**Runtime (h)**
Behavioral model	4	2,000	2,000	0.95	15	48	0.5-1
Full model (with calibration)	1	1,000	1,000	0.95	15	48	10-12

A.4 The potential of classifiers' confidence scores as covariates

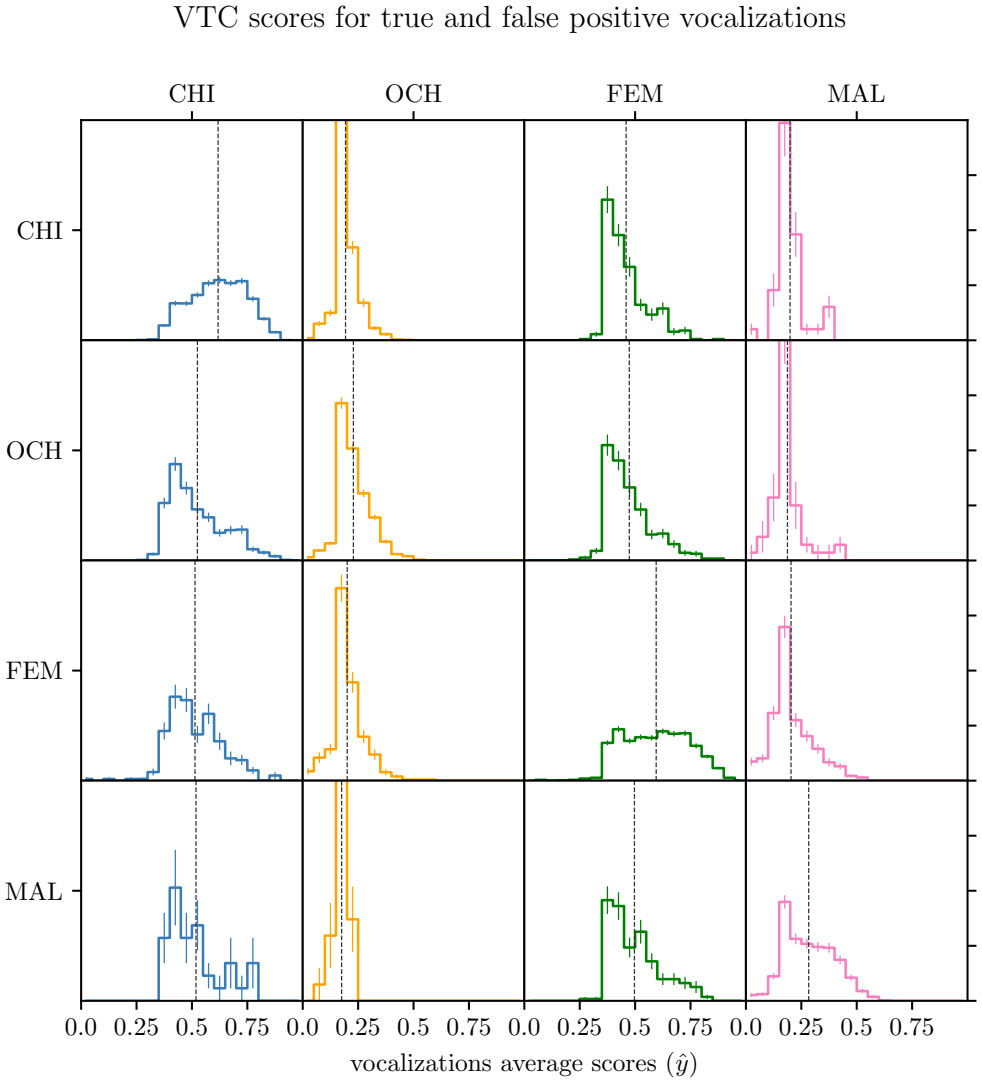


Figure 21: Distribution of the confidence score of the VTC for each vocalization given the detected speaker type (in columns) and the true speaker (in rows). Dashed vertical lines indicate the mean of the distribution. Vocalizations for which the speaker is correctly identified exhibit higher confidence scores, which suggests these scores could be used as informative covariates in a calibration approach.