INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Towards detection of canonical babbling by citizen scientists: Performance as a function of clip length

Amanda Seidl[1], *Anne S. Warlaumont*[2], *Alejandrina Cristia*[3]

[1]SLHS, Purdue University, West Lafayette, IN, USA
[2]University of California, Los Angeles, CA, USA
[3]Laboratoire de Sciences Cognitives et de Psycholinguistique, Département d'études cognitives, ENS, EHESS, CNRS, PSL University, Paris, France

aseidl@purdue.edu, warlaumont@ucla.edu, alecristia@gmail.com

Abstract

Theoretical, empirical, and intervention research requires access to a large, unbiased, annotated dataset of infant vocalizations for training speech technology to detect and differentiate consonant-vowel (canonical) syllables in infants' vocalizations from less mature vocalizations. Citizen scientists could help us to achieve the goal of this dataset, if classification is accurate regardless of coders' native language and training and can be completed on clips short enough to avoid revealing personal identifying information. Three groups of coders participated in an experiment: trained native, semi-trained native, and minimally-trained foreign. When vocalizations were presented whole, reliability was highest across the trained coders, with little difference between the semi-trained and minimally-trained coders. Among minimally-trained coders, reliability for 400ms-long clips was very similar to that found for full clips, with lower values for 200 and 600ms clips. Finally, error rates were minimized when 400ms-long clips were used. In sum, minimally-trained coders can achieve fairly reliable and accurate results, even when their native language does not match infants' target language and when provided with very short clips. Since shorter clips protect the identity of the child and her family, this manner of data annotation may provide us with a way of building a large, unbiased dataset of infant vocalizations.

Index Terms: Language acquisition, citizen science, large-scale annotation, infant vocalization, canonical babbling

1. Introduction

Childhood speech and language disorders carry important costs for both the individual and the society. For example, young children who have lower language skills are more likely to have difficulty with reading and writing tasks in school, more likely to encounter social challenges, and more likely to have poorer academic and professional outcomes later in life [1]. Thus, early diagnosis of risk for speech or language disorders is necessary to efficiently allocate interventions.

One promising area of research examines infants' vocal development around the emergence of canonical babbling (those repeated consonant-vowel sequences that infants begin produce in the middle of their first year, e.g., "ba ba ba"). A large body of recent work documents that the timing of the emergence of canonical babbling is predictive of later language outcomes. For example, in one study, infants who, at 10 months, had not reached the canonical babbling stage went on to have lower vocabulary size scores at 18, 24, and 36 months [2]. Similarly, in studies of children with higher risk factors for speech or language disorders (including genetic factors, e.g., Fragile X syndrome [3], and physical factors, e.g., hearing impairment [4]), canonical babbling emerges late and its emergence is linked to later language outcomes.

However, analyzing infants' vocal development is often done by highly trained professionals in limited environments (e.g., in a clinic room) and thus the cost is high and results may not be representative of each child's true performance. Specifically, many current analyses of vocal development across the first year of life are limited by short recording times (e.g., only 50 to 100 utterances analyzed at each age or vocal stage), circumscribed context (e.g., recordings are done in only the lab or only the home, limiting comparison across samples), small sample sizes (e.g., often only 20-30 children in even the larger studies), and homogeneous groups (e.g., there are few studies examining vocal development across a range of cultures, ethnicities, socio-economic status, bilingual status, etc.)

As a result, our current understanding of vocal development and its relationship to language outcomes is based on a narrow snapshot of early development – one which is unlikely to fully represent children who are developing atypically or typically. These barriers constrict efforts to determine when intervention is necessary accurately, and to measure effectiveness of interventions over time precisely.

Recent years have seen the rise of an alternative approach to describing early language development: Instead of observing the child for a short period of time in a very restricted situation, infants are fitted with a recording device over many hours or days, which allows researchers to sample the child's natural behavior in a wide range of contexts. While collecting these data is trivially easy, annotating them is a nearly insurmountable roadblock. To begin with, previous studies have relied on highly trained annotators, and many consider the training difficult. In addition, to apply this same approach to recordings lasting whole days, it would be necessary to comb through these large audio files to find child vocalizations and separate them from adult vocalizations and other noise, which is extremely time consuming. In fact, previous manual coding studies in one of our labs have yielded a coding rate of about 2 minutes of audio per hour. At this rate, a full-day of recordings of the child (typically 12 to 16 hours of recording time) would take a coder 360 to 480 hours – not including extra time needed to ensure coders are reliable with each other. This labor-intensive training and annotation process explains why it has been so far impossible to generalize previous methods to such large datasets.

A solution to this could be found by automatizing (or semi-automatizing) this process, for instance using a diarizer or speaker detection system to find child vocalizations, and a vocal maturity classifier to tag them as canonical or not; or per-

http://dx.doi.org/10.21437/Interspeech.2019-1773

haps using an end-to-end system which performs both tasks at once. The basis for such a technological solution is currently emerging: The LENA Foundation built a private large-scale, annotated dataset [5], on which they trained several classifiers [6]. One of them performs talker diarization, indicating where in the whole recording the child is vocalizing. This classifier can be bought for a fee, and has allowed data collection in many sites and many different populations [7]. Classifiers for vocal complexity are also being developed by the LENA Foundation [8]. However, like their dataset, their code is private and thus cannot be reused or improved upon. Moreover, the Foundation's focus is mainly on American English learners aged 0-3 years, which means that classifiers created may or may not generalize to other datasets. Similar issues affect the data that have been studied in the past, most of which is not shared other than as sample audios and supplementary materials, and nearly all of which has been gathered in a narrow set of child populations and very often in controlled acoustic conditions. As a result, we do not have enough data to train open source classifiers, and even if we try, given the data sets, they may be biased to specific child populations.

In an ongoing collaborative project, we set out to enable a larger cohort of child development researchers along with citizen scientists to contribute to the development of a large, multicultural, and unbiased dataset which could serve to train classifiers. Researchers who contribute data use the LENA automated talker diarization software (or open alternatives like DiViMe [9]) to find child vocalizations in long recording. This leaves the problem of annotating these vocalizations for vocal maturity, which, as noted above, is also extremely time-consuming. Our inspiration was to build on the citizen science movement [10]: Citizen scientists have successfully helped with a number of research projects, including identifying in an audio clip what talkers are eating while they speak (e.g., chips or pudding, [11]). However, automatically detected child vocalizations could contain identifying information from caregivers (e.g., in a false positive which includes a parent saying their credit card number), which we should not play over the web. We reasoned that a solution to this problem would be to take sound clips that are so short as to contain no more than one or two syllables. This would enable more researchers to contribute their recordings because it makes the data virtually unidentifiable. This is only a good move, however, if such short sequences can nonetheless allow accurate and reliable coding.

Thus, the primary aim of this paper is to report on an experiment aimed at assessing whether coders can accurately and reliably code short sequences of infant vocalizations, after short training. Specifically, we describe an experiment carried out to assess how reliable adults with varying levels of training are in detecting canonical syllables as a function of clip length (whole vocalization, as opposed to clips of 200, 400, or 600 ms in length).

2. Methods

2.1. Reference dataset

We built on an extant human-labeled dataset [12]. In that work, 16 North American English-learning infants were recorded longitudinally in a nursery-like laboratory during the period from 3 to 20 months of age. The infant vocalizations were first hand-tagged using a breath group criterion by members of Heather Ramsdell-Hudock's lab. Next, vocalizations were played in random order to an expert annotator and two trained undergraduate students, who counted the number of canonical syllables in each vocalization, the number of non-canonical syllables, and indicated whether the vocalization was likely a cry, laugh, or vegetative sound, whether it appeared to be mis-identified as an infant vocalization, or whether there was overlap from another sound source such as a toy or another human voice. Canonical syllables were defined as "adult-like syllables containing at least one consonant other than 'h' and at least one vowel". The two undergraduate students were trained by reading relevant literature, undergoing example-based training via the IVICT program [13], and discussing the concepts with the principal investigator (Anne Warlaumont). The number of canonical syllables per utterance and per syllable of any type increased significantly with age, with $r = .59$ and $r = .54$, respectively. Intercoder correlations in the whole dataset were strong, $rho = .74$, for number of canonical syllables in an utterance, providing a baseline against which automated methods were evaluated. Here we use this number as a baseline against which reliability can be gauged for citizen-scientists identifying canonical syllables within very short segments of infant utterances. That is, these judgments are treated as the "gold".

We selected vocalizations for which there were 2 coders, neither of whom had said the vocalization overlapped with other noise nor that the infant was crying or fussing. Vegetative and mislabeled units were also excluded. This resulted in 53 vocalizations being used for the present study.

Full vocalizations were cut into 200, 400, and 600 ms clips using a Praat script, available from the online supplementary materials https://osf.io/gq3jc/. For the 200 ms condition, this resulted in 1 clip for 12/53 vocalizations (that is, 12 vocalizations were already 200 ms long or shorter); for the remaining ones, one vocalization resulted in 2-15 clips, with a mean of 3.02 and a median of 2 (i.e., on average, vocalizations were about 600 ms long, resulting in 3 200 ms clips). For the 400 ms condition, this process resulted in 1 clip for 38/53 vocalizations; for the remaining ones, one vocalization resulted in 2-7 clips, with a mean of 1.47 and a median of 1. For the 600 ms condition, the cutting process resulted in 1 clip for 48/53 vocalizations; for the remaining ones, one vocalization resulted in 2-5 clips, with a mean of 1.15 and a median of 1. The beginning and end of each clip were amped from zero to full volume (or vice versa) in 10 ms units, to avoid clicks or other artifacts.

2.2. Participants

One group of participants were 12 American English-speaking undergraduate students from a communication sciences and disorders department. Semi-trained coders had been working in the lab for 1-2 years on various projects related to infant perceptual development and infant vocal development. They had all taken an introductory language development class and had ample experiences working with infants and children in their daily lives (e.g., babysitting, etc..). In addition, they viewed the same materials as the minimally trained group.

The second group were 18 people recruited through the RISC participant pool in France (https://expesciences.risc.cnrs.fr/). These individuals viewed an online presentation, and then completed a quick test to assess their ability to annotate vocal development. Both training and test are available online currently on https://purdue.ca1.qualtrics.com/jfe/form/SV_b1vSwOrTrbznYCV. We refer to this group as minimally trained.

2.3. Experimental design

The experiment was implemented in Open Sesame [14]. We used a fully crossed design: All participants heard all clips in 4 conditions: full, 200, 400, 600 ms; order was counterbalanced across participants within each group.

When full vocalizations were presented, the task was the same as for the gold annotators in terms of counting canonical syllables – with the exception that the gold coders were able to listen to the same clip several times before making a decision, whereas our coders were only allowed one playback (to keep the duration of the experiment short). Thus, in this condition, the outcome measure was, for each stimulus, the number of syllables the participant reported.

For the 200, 400, and 600 ms clips, the task was to answer, for each clip, whether there was a canonical transition or not. This allowed the task to be much simpler and faster (which is ideal for citizen science platforms), but it also meant that the participants did not provide a count. Instead, in these conditions, the number of syllables is calculated as the sum of "yes" judgments – that is, if a full vocalization was cut into three clips, and a participant answered "yes", "yes", and "no", then this clip would receive a count of 2 syllables for this participant.

3. Results

Data and software to reproduce the following results can be found on https://osf.io/gq3jc/. For an unknown reason, numeric judgments (i.e., for the full vocalization condition) were not captured by the software for three minimally-trained participants. Since comparing across conditions is crucial in this study, their data are altogether removed from analysis. Further, there are 42 judgments for counting canonical syllables that are removed because participants entered illegal characters (non-numeric, and not "i" either, which was the response for "junk"/not a child vocalization). Finally, we analyzed the distribution of "i" responses: If 200-600 ms is too short to process appropriately, then we may observe very high "i" judgments for the clips conditions. This was not the case: Whereas for the full condition 6% of the trials received an "i" judgment or an otherwise illegal response (as just mentioned), this affected 7-8% in the clips conditions. Notice that "i" judgments entail the exclusion of the judgment for the full vocalization trials, whereas for the clips conditions such a judgment simply does not count towards the total number of canonical syllables.

We used non-parametric (Spearman) correlations to compare the judgments by different types of coders. Note that in this analysis, one should concentrate on the estimate and not the p-value, since the latter will be affected by the number of data points, which is lower for the correlations between the two gold annotators (based on 53 full vocalizations) than among the gold annotators and either of the other groups (53 clips x number of participants). To compare correlations across groups, we employed confidence intervals, calculated via Fisher transformation.

For whole vocalizations, the reliability across the gold annotators was very high, above r>.9. The Spearman correlation estimates between the average of the judgments by these two trained annotators against the judgments by semi-trained and minimally-trained ones are shown on Figure 1. Notice that the confidence intervals never cross the topline, given by the agreement between two trained annotators, indicating a lower level of agreement with our coders; but they also never cross the random chance baseline. Most importantly for our purposes, one

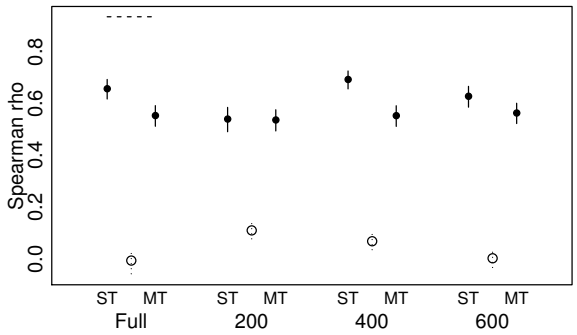

Figure 1: *Spearman correlation between collective judgments of semi-trained (ST) or minimally-trained (MT) annotators and the average of the gold annotators. The filled circle indicates the estimate, with vertical lines indicating the 95% confidence intervals (via Fisher transformation). The dotted lines at the top indicate the topline, i.e. the Spearman correlation estimate between the two gold annotators, which is only available for the full vocalizations condition. The open circles at the bottom show chance-level correlations (via resampling, N=50, dashed lines indicate 95% confidence intervals).*

of the clip conditions led to agreement that was as high as that found for the full vocalizations, namely the condition where vocalizations had been cut into 400ms-long clips (semi-trained: r=.66 full, r=.69 400ms; minimally trained: r=.55 full, r=.55 400ms). Furthermore, Figure 2 shows estimates for individuals within the semi-trained and minimally-trained groups against the average for the two gold coders. This shows that a majority of the minimally-trained individuals obtained correlations comparable to that of the semi-trained individuals.

Finally, we assessed whether syllable count error rates were similar across conditions. We focused on the data from minimally-trained participants, whose native language was not English, and we sub-sampled the data to a lower number of participants so as to assess the worst case situation. Results are shown in Figure 3. Interestingly, error rates were extremely stable even with as few as 5 participants. Negative errors for the 200 ms and full vocalization conditions indicates participants tended to report fewer syllable numbers than the referent annotators. It may be that the full vocalization condition tended to result in underestimates because of the binary nature of the task. On the other hand, the negative errors for the shortest 200 ms condition might be due to participants' missing canonical transitions. Positive error rates for the 400 and 600 ms conditions suggest that participants over-estimated the presence of canonical transitions in these cases, which may be due to artifacts of hearing a clip begin when the vocalization is already ongoing.

More relevant to our question is the absolute error rate, i.e. to what extent participants are wrong, above and beyond any systematic under- or over-estimation. Absolute error rates were minimized when 400ms-long clips were used at around .1, with 600 ms in second place at about .15, with .2 error rates for 200 ms and full vocalizations.

4. Discussion

We set out to assess to what extent naive individuals could provide reliable judgments of child babbling. The difference between semi-trained and trained individuals in the full vocaliza-

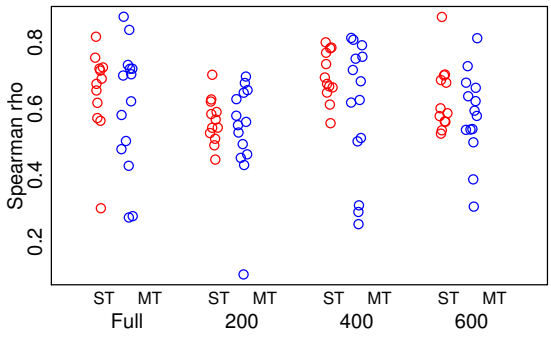

Figure 2: *Spearman correlation between judgments of individual semi-trained (blue) or minimally-trained (red) annotators and the average of the gold annotators. Each point indicates the estimate for one individual coder, in one specific condition. Points have been jittered to facilitate inspection.*

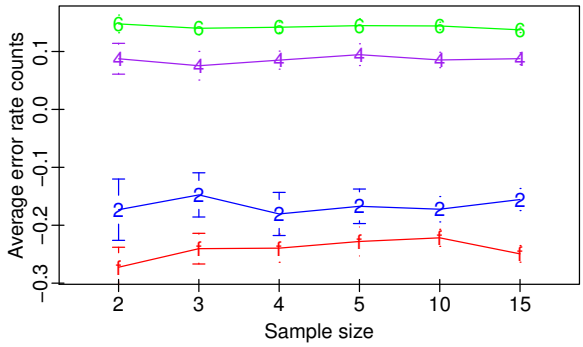

Figure 3: *Average error rate in canonical syllable counts for each condition against the average of the trained annotators. The red line shows data for the full vocalizations, blue for 200 ms, purple for 400 ms, and green for 600 ms. Error bars indicate 95% confidence intervals (via resampling, N=50).*

tions condition could be due to on or both of two key factors, which are confounded in the present data: (1) extensive training helps listeners detect canonical syllables more accurately; and (2) being able to listen to a vocalization several times may lead to more accurate counts than having to count on the fly after a single time of hearing a vocalization. The difference between semi- and minimally-trained individuals in the full vocalization condition (and presumably the other conditions) may also be due to several causes, since these groups differed on more than just the level of training. Indeed, native language background varied, and perhaps also the level of motivation to learn about child vocalizations may have led the semi-trained to be more attentive and careful than the minimally-trained individuals. In any case, it appears that totally naive participants can be quickly trained to achieve a respectable correlation, with some able to score comparably to individuals with much higher levels of exposure to infants and knowledge of infant development.

Our second goal was to assess whether full vocalizations were required, or shorter clips could be presented. Results suggest that performance for full vocalizations and 400 ms clips was comparable, with lower scores for 200 and 600 ms-long clips. For the 600 ms condition, this lower performance may be due to the task, because the participant could not say "yes" several times, leading to a ceiling effect on the reported counts. As mentioned previously, most of the original clips were in fact 600 ms or less, and thus however many syllables were originally in them, the participants could only provide a 0 or 1 count. The same argument cannot explain the relatively lower performance in the 200 ms condition. In this case, the problem may arise from a limitation in the participants' speed of processing the information (although note that this condition did not lead to a great deal more of "junk" judgments), or else on the availability of sufficient acoustic information. One aspect that we did not manipulate here was whether syllables were appropriately cut, e.g. at vowel middles or perhaps at consonantal edges. Although this may be theoretically interesting, we do not think this would be desirable as it would imply that, for the analysis of large data sets, one would have to be able to accurately detect vowel middles and/or consonantal edges. If such a classifier existed, then it could be used to detect canonical syllables to begin with.

An incidental finding in this project was that if we had a perfect vocalization activity detection, presenting only 400ms clips would result in 70% of the clips being presented whole, because at least in this corpus infant vocalizations were very short. Unfortunately, present-day voice activity detectors (VAD) are far from being perfect, particularly when applied to child speech. Anecdotal reports from the DIHARD Challenge [15, 16], which contains child data, suggests that VADs tend to miss infant vocalizations, with recall rates as low as 30%. Moreover, we would also need excellent precision to avoid including adults' speech, which may contain personally identifying information. Although systematic data on this is also missing, one evaluation of the LENA system found that 15% of vocalizations were tagged as child when they were in fact by an adult or vice versa in [17]. Thus, at the time being, we cannot suppose that diarizer will be perfect, and thus cutting vocalizations into shorter clips may still be desirable.

In short, we found that minimally trained coders can achieve fairly reliable and accurate results, even in a case of mismatching native language and when provided with short extracts. Since these extracts protect the identity of the child and her family, this manner of data annotation may finally provide us with a way of building a large, unbiased dataset of infant vocalizations on which to train classifiers. Indeed, informed by the results of this study, a collaborative project among members of DARCLE (darcle.org) led to the creation of a Babble corpus used in this year's ComParE challenge [18], in which short clips were annotated via a citizen science platform [11]. We hope this report inspires others to attempt similar endeavors.

5. Acknowledgements

We thank the other members of the BabbleCor collaborative project (Gladys Baudet, Elika Bergelson, Marisa Casillas, Meg Cychosz, Mark Liberman, Camila Scaff, Lisa Yankowitz, and especially Heather Ramsdell-Hudock, who shared with us the original data which allowed the present study). AC acknowledges the support from Agence Nationale de la Recherche (ANR-17-CE28-0007, ANR-16-DATA-0004 ACLEW, ANR-14-CE30-0003 MechELex, ANR-17-EURE-0017); and the J. S. McDonnell Foundation Understanding Human Cognition Scholar Award. ASW acknowledges support from the National Science Foundation through grant numbers BCS-1529127 and SMA-1539129/1827744 and from a James S. McDonnell Foundation Understanding Human Cognition Scholar Award.

6. References

[1] B. Hassinger-Das, T. S. Toub, K. Hirsh-Pasek, and R. M. Golinkoff, "A matter of principle: Applying language science to the classroom and beyond." *Translational Issues in Psychological Science*, vol. 3, no. 1, p. 5, 2017.

[2] D. Oller, R. Eilers, A. Neal, and A. Cobo-Lewis, "Late onset canonical babbling: A possible early marker of abnormal development," *American Journal on Mental Retardation*, vol. 103, no. 3, pp. 249–263, 1998.

[3] K. Belardi, L. Watson, R. Faldowski, H. Hazlett, E. Crais, G. Baranek, and D. Oller, "A retrospective video analysis of canonical babbling and volubility in infants with Fragile X syndrome at 9– 12 months of age," *Journal of Autism and Developmental Disorders*, vol. 47, no. 4, pp. 1193–1206, 2017.

[4] S. Ilyer and D. Oller, "Prelinguistic vocal development in infants with typical hearing and infants with severe-to-profound hearing loss," *Volta Review*, vol. 108, no. 2, pp. 115–138, 2008.

[5] J. Gilkerson and J. A. Richards, "The LENA Natural Language Study," 2008, lENA Technical Report (LTR-02-2). Boulder, CO: LENA Foundation.

[6] D. D. Xu and T. D. Paul, "System and method for expressive language, developmental disorder, and emotion assessment," 2014, uS Patent App. 14/265,188.

[7] M. Vandam, A. S. Warlaumont, E. Bergelson, A. Cristia, M. Soderstrom, P. De Palma, and B. Macwhinney, "Homebank: An online repository of daylong child-centered audio recordings." *Seminars in speech and language*, vol. 37, no. 2, pp. 128–142, 2016.

[8] D. K. Oller, P. Niyogi, S. Gray, J. Richards, J. Gilkerson, D. Xu, U. Yapanel, and S. Warren, "Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development," *Proceedings of the National Academy of Sciences*, vol. 107, no. 30, pp. 13 354–13 359, 2010.

[9] A. Le Franc, E. Riebling, J. Karadayi, Y. Wang, C. Scaff, F. Metze, and A. Cristia, "The ACLEW DiViMe: An easy-to-use diarization tool," in *Proceedings of Interspeech*, 2018.

[10] H. Dickinson, L. Fortson, C. Scarlata, M. Beck, and M. Walmsley, "Modeling with the crowd: Optimizing the human-machine partnership with zooniverse," *arXiv preprint arXiv:1903.07776*, 2019.

[11] S. Hantke, F. Eyben, T. Appel, and B. Schuller, "ihearu-play: Introducing a game for crowdsourced data collection for affective computing," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 891–897.

[12] A. S. Warlaumont and H. L. Ramsdell-Hudock, "Detection of total syllables and canonical syllables in infant vocalizations." in *INTERSPEECH*, 2016, pp. 2676–2680.

[13] K. Oller, E. Buder, A. Warlaumont, R. Dale, B. Franklin, Y. Jhang, C.-C. Lee, N. Rangisetty, E. Bene, and L. mei Chen, "IVICT (Infant Vocalization Interactive Coding Trainer)," 2015, www.babyvoc.org/IVICT.html.

[14] S. Mathôt, D. Schreij, and J. Theeuwes, "Opensesame: An open-source, graphical experiment builder for the social sciences," *Behavior research methods*, vol. 44, no. 2, pp. 314–324, 2012.

[15] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "First DIHARD challenge evaluation plan," 2018, https://zenodo.org/record/1199638.

[16] ——, "The first DIHARD speech diarization challenge," https://coml.lscp.ens.fr/dihard/2018, accessed: 2019-03-17.

[17] A. Seidl, A. Cristia, M. Soderstrom, E.-S. Ko, E. A. Abel, A. Kellerman, and A. Schwichtenberg, "Infant–mother acoustic–prosodic alignment and developmental risk," *Journal of Speech, Language, and Hearing Research*, vol. 61, no. 6, pp. 1369–1380, 2018.

[18] B. W. Schuller, A. Batliner, C. Bergler, F. B. Pokorny, J. Krajewski, M. Cychosz, R. Vollmann, S.-D. Roelen, S. Schnieder, E. Bergelson, A. Cristia, A. Seidl, A. Warlaumont, L. Yankowitz, E. Nth, S. Amiriparian, S. Hantke, and M. Schmitt, "The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity," in *Proceedings of Interspeech*, submitted.