The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages
29-31 August 2018, Gurugram, India

Diarization in Maximally Ecological Recordings: Data from Tsimane Children

Julien Karadayi[1,2], *Camila Scaff*[1], *Jonathan Stieglitz*[3], *Alejandrina Cristia*[1]

[1] LSCP, Département d'études cognitives, ENS, EHESS, CNRS, PSL University, Paris, France
[2] Institut national de recherche en informatique et en automatique, Paris, France
[3] Universite Toulouse 1 Capitole, Toulouse, France

alecristia@gmail.com

Abstract

Daylong recordings may be the most naturalistic and least invasive way to collect speech data, sampling all potential language use contexts, with a device that is unobtrusive enough to have little effect on people's behaviors. As a result, this technology is relevant for studying diverse languages, including understudied languages in remote settings – provided we can apply effective unsupervised analyses procedures. In this paper, we analyze in detail results from applying an open source package (DiViMe) and a proprietary alternative (LENA^{TM}), onto clips periodically sampled from daylong recorders worn by Tsimane children of the Bolivian Amazon (age range: 6-68 months; recording time/child range: 4-22h). Detailed analyses showed the open source package fared no worse than the proprietary alternative. However, performance was overall rather dismal. We suggest promising directions for improvements based on analyses of variation in performance within our corpus.

Index Terms: ecological recordings, language acquisition, voice activity detection, speaker diarization

1. Introduction

With recent advances in voice recording technology, it is now easy to collect large amounts of speech data using a small, lightweight recording device worn on one's shirt. This type of technology may be ideal to document languages spoken in societies that are changing fast, as they require nothing of the person wearing the recorder, and they are often unobtrusive enough to allow research participants to feel unobserved and thus act naturally. Despite these and other advantages, daylong recordings pose a number of challenges, and it is unclear how broadly available speech technology fares with these data. In this paper, we present a detailed analysis of unsupervised talker diarization results (i.e., automatic labeling of 'who spoke when') obtained using the open source package DiViMe [1] on data collected from a community speaking an under-resourced language, Tsimane.

1.1. Why are daylong recordings interesting?

We focus on the case of language development both because our data are drawn from this domain, and because such data clearly shows the many ways in which daylong recordings are technically challenging. As the child and those surrounding him/her go about their typical day, the acoustic environment for the conversations varies: a family may start the day in a quiet bedroom, have meals in a kitchen where a radio is on, and later have an open-air stroll to visit a neighbor while birds are chirping. In the case of multilingual societies, the language spoken may change within individuals across settings - or the presence and frequency of "code switching" may change. Additionally, the resulting recordings may contain speech that is unfamiliar to usual processing models: Early on, the majority of children's vocalizations will be non-speech (e.g. cooing or crying) but nonetheless trigger communicative reactions from social partners, so that a model that excludes non-speech vocalizations would fail to detect the extant structure in these cases. As children begin speaking with age, their pronunciations may be deviant and speakers may not respect turn-taking. Furthermore, previous targeted research (e.g., [2]) shows that adults in these conversations with children do not behave in the same way as they do when talking to other adults. Specifically, adults' speech when talking to infants has larger pitch excursions (deviating from the adult-adult norm acoustically) and utterances tend to be shorter.

1.2. How does available software fare on daylong recordings?

Although there exist some state-of-the-art solutions that have been developed for extensive, English-spoken datasets (e.g., [3]), the relevant software is not yet readily available, and thus cannot be viewed today as potentially re-trainable and generalizable for other languages and populations. There is one exception, which concerns a proprietary set of products. The LENA^{TM} Foundation developed easy-to-use, commercially available hardware (300 US$ per unit) and software (about 12k US$ per license, including an elegant graphical user interface). This enabled developmental psychologists and speech language pathologists to use daylong recordings, without requiring any technical expertise, to the point that at present basically all developmental studies using daylong recordings actually employ LENA^{TM} technology.

The LENA^{TM} software was introduced and explained in a number of publications by the Foundation and associated researchers [4, 5, 6, 7]. We are not aware of specific reports by the LENA^{TM} Foundation evaluating talker diarization using typical metrics found in the speech research community (such as diarization error rate or frame-wise accuracy). Instead, accuracy is more often reported in terms of agreement, which is 71-86% when diarization labels are collapsed into the following broad categories: "adult", "child", "TV", and "other" ([8], see similar performance reported in [9, 10]).

LENA^{TM} products have since been used by researchers seeking to document language acquisition in French, Swedish, Dutch, Spanish, and Mandarin Chinese. Given that the LENA^{TM} software is proprietary, no adaptation or re-training was performed. Nonetheless, researchers did try to assess accuracy for their language. However, evaluation for talker diarization performance specifically also tends to be limited, as researchers' interest typically focus on global language input metrics, such as rough estimations of the number of adult words or syllables recorded in a day [11, 12, 13, 14, 15].

10.21437/SLTU.2018-7

Fortunately, a new conversational analysis suite that is open source is currently being developed [1]. The Diarization Virtual MachinE (DiViMe for short) currently contains two tools that permit speech activity detection (i.e., detecting which portions of the recording contain some speech), and one tool for speaker diarization (i.e., attributing a speech portion to one or another speaker), which, combined, lead to two purely unsupervised pipelines yielding a segmentation of the recording into different speakers. In a recently accepted paper [16], global speech activity detection and talker diarization performance was reported.

1.3. Present work

In the present paper, we sought to analyze in more detail the performance of the DiViMe tools, comparing when possible against the LENA^{TM} software. We asked the following specific questions:

1. Do tools within DiViMe fall below, at, or above the level of accuracy of the LENA^{TM} software? Given that the LENA^{TM} software has been developed using the same kind of developmental speech data we use here, we predicted DiViMe's accuracy would be significantly lower than that of LENA^{TM}
2. Is performance affected by using the LENA^{TM} hardware versus alternative hardware? Based on the same within-domain reasoning, we expected LENA^{TM} software to perform better on recordings made with the LENA^{TM} hardware than with alternative recorders. We expected no difference for the DiViMe tools applied to LENA^{TM} or non-LENA^{TM} hardware.
3. How is performance affected by factors varying across clips, such as the amount of speech found in a clip? We made no precise predictions for this last, more exploratory, section.

As will become obvious, in the process of this analysis, we also encountered some methodological and conceptual issues related to evaluation that become salient in the context of highly ecological recordings.

2. Corpus

2.1. Population and language

The Tsimane (also called Chimani, Chimane, or Tsimané) are a forager-farmer, mostly monolingual, indigenous population living in $>$ 85 relatively small villages (50-500 residents/village) in lowland Bolivia. The Tsimane language is an isolate; there is some disagreement as to whether Mosetén and Tsimane should be considered two languages in the same family, or merely regional variants of a single language (e.g., [17] states the latter). Tsimane phonology was most recently described in the grammar by the missionary Wayne Gill [18], although records in other resources (e.g., [19]) vary in terms of the precise number and identity of sounds in the phonological inventory (possibly due to the disagreements regarding whether Tsimane and Mosetén inventories are one and the same language). Nonetheless, it appears that there are at least 10 vowels, with 5 height-frontness qualities fully crossed with nasalization; and at least 17 consonants, including fricatives, as well as plain, nasal, affricated, and aspirated stops, and affricated-aspirated stops.

Previous work shows that Tsimane children are cared for and talked to mainly by their mother, although many others (both children and adult) provide some physical care and some speech [20]. Due to high fertility, Tsimane households are large relative to typical American households; for instance, in the present sample, households tended to contain 4 to 5 siblings in addition to 2 or more adults. Additionally, children can and do walk around beyond their house, and thus the number of unique voices that may be registered in the device is, maximally, the village size (about 200).

2.2. Recordings

The data used here come from [21]. Recordings were collected by CS and a Tsimane research assistant in July 2017 in one village. A total of 15 families agreed for their child(ren) to wear a recording device, and the entire village agreed to participating in the study.

Participating families were visited one by one. If the family had several children within the sampled age range (6 months to 68 months), all children were equipped with a recording device at the same time. A total of 27 different children (9 female) were equipped with a recording device on a first visit, including 9 pairs of siblings, one set of triplets, and four children who did not have siblings that fell within the sampled age range. Of the 27 children, 14 agreed to the second recording, for a total of 41 recordings. Recordings were started either at around 9 AM (N=25), and at around 6pm (N=16). Two of the recordings have not yet been coded, resulting in a final recording N=39, and a final child N=26. We aimed to record for 12-16 hours, and 27 recordings were this long. In 3 cases of device malfunction, as few as 4-5h were recorded; the rest of the recordings were between 8 and 22h length. Notice that, when the device is turned on in the evening, much of this recording covers night time.

We intended to use three different brands of recording devices, both to maximize the number of devices delivered simultaneously, and to check to what extent SAD and diarization performance is affected by the recording device (or, put otherwise, to what extent we can generalize across devices). We took to the field one Olympus© WS831 (80 each), 4 Etekcity© USB recorders (12 each), and 4 LENA^{TM} devices (300US$ each). While in the field, we found that families tended to disprefer the Olympus due to its weight and relative bulkiness, resulting in the following distribution of recordings: 3 made with the Olympus; 20 with USBs; and 18 with LENA^{TM} . In all cases, children wore a custom-made t-shirt with a breast pocket in which a recording device fit snugly. We used LENA^{TM} clothing for LENA^{TM} devices, and either a contractor or CS sowed pockets on commercially-sold t-shirts for the other two devices.

2.3. Annotations

For each recording, we extracted one minute per hour (on average 13 annotated minutes, ranging from 4 to 16 minutes). In all, the present analyses are based on 537 one-minute extracts. A trained phonetician labeled each minute into segments of vocalizations using Praat software [22]. See Figure 1 for an example of the labels used on this corpus.

Vocalizations were classified into: the focal child, the "main female voice" (MFV, usually the mother), other female adults, other male adults, and other children. For the experiments presented here, only this "pure" speech was considered.

In the interest of completeness, we point out that there were a few more tiers annotated. Specifically, if more than 2 people were talking at the same time unintelligibly, then the coder marked this segment as "2 or more speakers talking at the same time". Also, if speech sounded far away to the point that it was

difficult to tell who was talking or when, these vocalizations were annotated in a "background speech" tier. There was also a tier for music or radio (labeled as "Noise"), used only if it overlaps with actual speech. In future work, we should explore how results are affected by the inclusion of the talker overlap and background speech tiers.

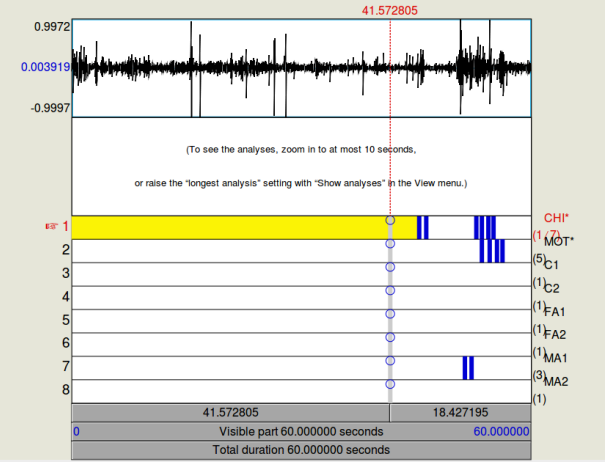

Figure 1: *Screenshot of the annotation using Praat. The labels used here are CHI for the focal child, MOT for the "main female voice", C for other children, FA for other female adults and MA for other male adults.*

2.4. LENA annotations

Since about half of the recordings were gathered with LENA^{TM} hardware, in these analyses we could include 272 1-minute extracts which could be analyzed with the LENA^{TM} software. Note that one can only use their software if recordings are gathered with their hardware. Therefore, the remaining 265 clips cannot be used towards assessing the performance for LENA^{TM} .

We describe briefly the LENA^{TM} pipeline, following [4, 5, 6, 7]. Using the full recording as input, the LENA^{TM} software extracts 36 mel frequency cepstral coefficients and their deltas in 25 ms windows every 10ms. These features are then analyzed with a system performing joint segmentation and classification using a Minimum Duration Gaussian Mixture Model (MD-GMM), where the minimum duration is .6s for most classes (further using dynamic programming to find the sequences with maximum likelihood). The MD-GMM model was trained on over 150 hours (extracted from daylong recorders worn by American English-learning children aged 0-4 years, equal numbers of boys and girls), which were hand-segmented by professional transcribers and classified into the 4 categories which count towards Speech: Key Child (the one wearing the recording device), Other Children, Adult Males, and Adult Females; and the following non-speech categories: TV & other electronic sound, Noise, Silence, and Overlapping sound (overlap of any two categories, e.g., Key Child + Noise). When the system is uncertain (i.e., the best fitting category nonetheless has a low likelihood), then the segment is labeled "far" (e.g., Key Child Far). In pilot analyses we found better SAD performance without these "far" items. Therefore, in this study, we declared them as non-speech.

For the present study, a custom-written set of scripts were used to extract the one-minute sections corresponding to the analyzed clips from the LENA^{TM} -produced transcript of the whole recording. We generated two annotations, one only containing speech/non-speech; and the other maintaining the 4 speech classes provided by LENA^{TM} .

3. Tools

We generated two unsupervised talker diarization pipelines building on the tools currently included in DiViMe [1]. The two pipelines differ on which tool is used for the speech activity detection phase (either LDC-SAD or Noisemes); both pipelines use DiarTK for the talker diarization phase. See Figure 2 for a flow graph explaining the system.

3.1. Speech Activity Detection

The DiViMe system includes two options for Speech Activity Detection (SAD). The first one, called **LDC SAD**, applies a band-pass filter and extracts PLP (Perceptual Linear Prediction) cepstral coefficient using HTK, before using a broad phonetic class recognizer based on a GMM-HMM model. This system was trained on the Buckeye Corpus [23], which consists of American talkers being interviewed in a quiet room.

The second one, called **Noiseme SAD** [24], relies on a neural network to predict frame-level posteriograms of 17 types of sound events, also called "noisemes" [25]. These classes include speech-English, speech-non English, engine, cheers, crowd etc... The network architecture is a single bidirectional LSTM layer with 400 hidden units in each direction. It was programmed using the Theano toolkit and trained on 10 hours of HAVIC data [26]. The inputs of the network are 6,669 low-level acoustic features, reduced to 50 dimensions by a PCA, extracted with the OpenSMILE toolkit [27]. In the current DiViMe release, the "speech-English" and "speech-non English" classes are merged into a "speech" one, by summing their posteriorgrams, all the others are merged into a "non-speech" class, and each frame is attributed to the class with the highest probability.

3.2. Diarization

The DiViMe system contains one Talker Diarization (TD) tool called **DiarTK** [28]. This toolkit, written in c++ and open source, extracts MFCC features and performs a non-parametric clustering of the frames, using agglomerative information bottleneck clustering [29]. Each cluster of frames is then assimilated to one talker, and the most likely diarization sequence is extracted with a Viterbi realignement.

3.3. Evaluation

Evaluation tools are also included in DiViMe, for both the SAD and the diarization. For the SAD task, we use the evaluation included in the LDC SAD [30], which computes the false alarm (FA) rate (proportion of frame falsely attributed to speech by the system), and a miss rate (proportions of actual frames of speech that were attributed to non-speech by the system). For the TD task, we evaluate the systems with the Diarization Error Rate (DER), with the same definition as in the DiHARD challenge [31]. This error rate is the the sum of a false alarm rate, a missed speech rate, and a mismatch rate (proportion of frames attributed to the wrong speaker ID).

Given the way our data is sampled, we must define evaluation strategies when the studied audio recordings contain no speech at all, or when one of the two tasks returns no speech

attribution. Such a case is seldom discussed in the literature because, most often the studied recording are chosen at moments when there is speech. This is not the case in naturalistic recordings however, for which we should therefore adopt a coherent framework to treat every possible case.

For the SAD task, when the gold annotation is empty, if the SAD system returns no speech label then FA = 0 and Miss = 0; else, if the SAD system finds some speech, then FA = 100 and Miss = 0. When the gold annotation is not empty, if the system did not find any speech segment, then FA = 0 and Miss = 100.

The same rules are used for the talker diarization evaluation. Additionally, in each of the previous cases, a Mismatch = 0 is set, because one cannot misattribute a class to a frame if there are no corresponding classes in the gold, and one cannot misattribute a class when none has been detected either.

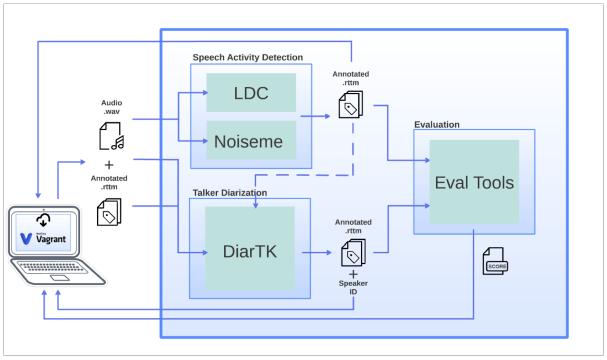

Figure 2: *Flow Graph of the tools in DiViMe.*

4. Results and discussion

There are 263 out of 537 files with no speech. Given that results could change when considering all files versus only files where there is at least some speech, we report statistics on both. Additionally, we observed 132 aberrant DER values (above 300% and up to 3,500%), of which most were found in the LDC (52) or LENA™ (62) pipelines (18 such values for noisemes). These values were declared missing so that they would not bias results, but it remains for future work to understand how they emerge, and whether alternative metrics need to be used for tasks like the present.

To focus on stable effects, we fit a mixed linear regression on each performance variable (M, FA, DER), with system, corpus, and their interaction as fixed effects, and clip identity as random effect. We declared the LENA™ software and the recordings gathered with LENA™ as baselines for system and corpus respectively. Results are shown on Table 1. These analyses suggest that LDC and Noisemes led to significantly higher miss rates than LENA™ , but the three systems did not differ significantly on false alarms and DER. Thus, it appears that the current open source alternative to LENA™ is competitive in terms of overall performance.

Our conclusions are less cheerful when we consider the overall levels of performance achieved. The LENA™ system resulted in an average DER of 121%; performance using the DiViMe pipelines on both LENA™ and NL clips led to DERs

Factor	M-all	M-SS	FA-all	FA-SS	DER-all	DER-SS
Intercept	1 (1)	3 (4)	76 (2)*	76 (2)*	130 (3)*	78 (10)*
LDC	7 (1)*	13 (5)	-13 (2)	-13 (2)	-5 (3)	11 (12)
Noi	33 (1)*	76 (6)*	-59 (2)	-59 (2)	-23 (3)	20 (12)

Table 1: *Results of mixed models fit to predict miss rate (M), false alarm rate (R) and Diarization Error Rate (DER) in all the files (-all) or the subset containing speech (-SS). Since all these are error rates, lower is better. In the rows, Noi stands for noisemes. Corpus and interaction effects are not shown.*

of 110%, which is still a great deal higher than performances reported elsewhere (e.g., [32]). To a certain extent, the difference is due to the fact that the systems in DiViMe are not state of the art. Indeed, the two systems included in DiViMe performed near the bottom in a recent challenge called DiHARD, aimed at assessing diarization performance in "difficult" datasets, such as meetings and doctor-patient interviews [33]. But it is likely that the present recordings are considerably more difficult even than their "difficult" datasets, since (according to [16]) the DiViMe pipelines applied to the DiHARD standardized evaluation set led to DERs of 65-72%. The difference between these DiHARD scores in the 70's and the DERs averaging 110% obtained for the Tsimane dataset represents the additional difficulty posed by these data.

What makes our data so difficult? Further analyses showed that this was not due to overlap, which was nearly 1% (excluding files with no speech), potentially due to the fact that speech that was difficult for our human annotator to diarize was classified in a different tier, and excluded from the present analyses. There remained several alternative hypotheses why our data are difficult. Considering only clips with speech, there is not a lot of speech overall (the average number of frames with speech for morning-delivered devices 13%, for evening-delivered devices it is 6.5%); turns are rather short (median turn duration is 1.2s, range 0.3-53s); and much of the speech comes from children, whose voices may not be well-classified by the tools in DiViMe without further retraining (on median, 42% of annotated speech consisted of child vocalizations; range over clips with speech 0-100

To check whether these factors explained variance across different clips, we fit two regressions predicting DER. In the simplest one, we declare device (whether the recordings were gathered with the LENA™ hardware or not), system (LDC, Noisemes), and their interaction as predictors, and clip as repeated measure.[1] The second regression added the following predictors: time of delivery (morning, evening), total amount of speech in the clip, proportion of the speech in the clip spoken by children as opposed to adults, average turn duration, and child age. Model comparison showed that these factors accounted for additional variance above and beyond device and system [$F(4,539)=10.4$, $p < .001$], although both the simple and the complex models actually accounted for small proportions of variance overall (adjusted $R^2 = 2\%$ and 9% respectively). Average turn duration was not in fact a significant predictor and time of delivery was marginal. As predicted, greater speech amount led significant reductions in DER (β=-5, SE 1), whereas the proportion of child speech increased DER (β=-95, SE 31).

[1] We preferred a simple over a mixed regression so as to be able to report proportion of variance explained; this meant that we could not include the LENA™ software as a system because non-LENA™ clips cannot be analyzed with the LENA™ software and simple regressions do not accommodate for missing cells. Also, only clips with some speech were included.

Thus, poor performance and variation across clips may be partially due to the low prevalence of speech overall, and the high proportion of children's speech found in these child-centered recordings, which clearly deviate from the recordings that have been the focus of interest in the past. A recent paper focusing on the talker diarization phase (i.e., using speech activity detection from human annotations) found that i-vectors retrained with a variety of materials including children's voices surprisingly failed to find improvements [34]. We hope others may continue to explore targeted retraining.

Overall, our regressions explained little variation. We hypothesize that uncoded variability in the characteristics of the acoustic environment, which may change greatly throughout a typical day, explains much variation in performance. We look forward in the future to incorporating speech enhancement techniques that could directly address this issue.

5. Conclusions

Detailed analyses revealed that the open source tools packaged within DiViMe were competitive against the current field standards, the LENA^{TM} software, at least for our dataset. However, performance was tremendously variable across clips and overall rather dismal, and much lower than that found even for the "difficult" clips used in the recent DiHARD Challenge [33]. We anticipate that improving performance will require direct attention to the unique challenges posed by these recordings, including variable speech enhancement techniques and retraining for the types and range of voices found here.

6. Acknowledgements

This work was supported Agence Nationale de la Recherche (ANR-14-CE30-0003 MechELex, ANR-10-IDEX-0001-02 PSL*, ANR-10-LABX-0087 IEC; and ANR-16-DATA-0004 ACLEW via the Trans-Atlantic Platform "Digging into Data" collaboration grant ACLEW: Analyzing Child Language Experiences Around The World), as well as funding from the J. S. McDonnell Foundation. We thank Adrien Le Franc for help with analyses and Mustapha Mardi for the annotations. We also directly benefited from interactions at the 2017 Frederick Jelinek Memorial Summer Workshop, which was supported by Amazon, Apple, Facebook, Google, and Microsoft (in alphabetical order); and used the Bridges system, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC); and the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575.

7. References

[1] "ACLEW diarization virtual machine," https://github.com/aclew/DiViMe.

[2] A. Cristia, "Input to language: The phonetics and perception of infant-directed speech," *Language and Linguistics Compass*, vol. 7, no. 3, pp. 157–170, 2013.

[3] C. Yu and J. H. Hansen, "Active learning based constrained clustering for speaker diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2188–2198, 2017.

[4] J. Gilkerson and J. A. Richards, "The lena natural language study," 2008.

[5] D. Xu, U. Yapanel, S. Gray, and C. T. Baer, "The LENA©language environment analysis system: the interpreted time segments (its) file," 2008.

[6] D. Xu, U. Yapanel, S. Gray, J. Gilkerson, J. Richards, and J. Hansen, "Signal processing for young child speech language development," in *First Workshop on Child, Computer and Interaction*, 2008.

[7] D. Xu, J. Gilkerson, J. Richards, U. Yapanel, and S. Gray, "Child vocalization composition as discriminant information for automatic autism detection," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*. IEEE, 2009, pp. 2518–2522.

[8] D. Xu, U. Yapanel, and S. Gray, "Reliability of the LENA©language environment analysis system in young childrens natural home environment," 2008.

[9] E.-S. Ko, A. Seidl, A. Cristia, M. Reimchen, and M. Soderstrom, "Entrainment of prosody in the interaction of mothers with their young children," *Journal of child language*, vol. 43, no. 2, pp. 284–309, 2016.

[10] M. VanDam and N. H. Silbert, "Fidelity of automatic speech processing for adult and child talker classifications," *PloS one*, vol. 11, no. 8, p. e0160588, 2016.

[11] A. Weisleder and A. Fernald, "Talking to children matters: Early language experience strengthens processing and builds vocabulary," *Psychological science*, vol. 24, no. 11, pp. 2143–2152, 2013.

[12] J. Gilkerson, Y. Zhang, D. Xu, J. A. Richards, X. Xu, F. Jiang, J. Harnsberger, and K. Topping, "Evaluating language environment analysis system performance for chinese: A pilot study in shanghai," *Journal of Speech, Language, and Hearing Research*, vol. 58, no. 2, pp. 445–452, 2015.

[13] M. Canault, M.-T. Le Normand, S. Foudil, N. Loundon, and H. Thai-Van, "Reliability of the language environment analysis system (lena) in european french," *Behavior research methods*, vol. 48, no. 3, pp. 1109–1124, 2016.

[14] I.-C. Schwarz, N. Botros, A. Lord, A. Marcusson, H. Tidelius, and E. Marklund, "The lenatm system applied to swedish: Reliability of the adult word count estimate," in *Interspeech 2017*. The International Speech Communication Association (ISCA), 2017, pp. 2088–2092.

[15] T. Busch, A. Sangen, F. Vanpoucke, and A. van Wieringen, "Correlation and agreement between language environment analysis (lena) and manual transcription for dutch natural language recordings," *Behavior research methods*, pp. 1–12, 2017.

[16] A. L. Franc, E. Riebling, J. Karadayi, Y. Wang, C. Scaff, F. Metze, and A. Cristia, "The aclew divime: An easy-to-use diarization tool," *Proceedings of Interspeech*, 2018.

[17] J. Sakel, *A grammar of Mosetén*. Walter de Gruyter, 2004, vol. 33.

[18] W. Gill, "A pedagogical grammar of the chimane (tsimane) language," *Ms., New*, 1999.

[19] S. Moran, D. McCloy, and R. Wright, Eds., *PHOIBLE Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology, 2014. [Online]. Available: http://phoible.org/

[20] A. Cristia, E. Dupoux, M. Gurven, and J. Stieglitz, "Child-directed speech is infrequent in a forager-farmer population: A time allocation study," *Child development*, 2017.

[21] C. Scaff, J. Stieglitz, and A. Cristia, "Daylong recordings from young children learning Tsimane in Bolivia," https://nyu.databrary.org/volume/445, accessed: 2018-03-03.

[22] P. Boersma, "Praat: doing phonetics by computer," *http://www.praat. org/*, 2006.

[23] M. A. Pitt, K. Johnson, E. Hume, S. Kiesling, and W. Raymond, "The buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability," *Speech Communication*, vol. 45, no. 1, pp. 89–95, 2005.

[24] Y. Wang and F. Metze, "A first attempt at polyphonic sound event detection using connectionist temporal classification," in *Proc. ICASSP*. New Orleans, LA; U.S.A.: IEEE, Mar. 2017.

[25] S. Burger, Q. Jin, P. F. Schulam, and F. Metze, "Noisemes: Manual annotation of environmental noise in audio streams," Carnegie Mellon University, Pittsburgh, PA; U.S.A., Tech. Rep. CMU-LTI-12-07, 2012.

[26] S. Strassel, A. Morris, J. G. Fiscus, C. Caruso, H. Lee, P. D. Over, J. Fiumara, B. L. Shaw, B. Antonishek, and M. Michel, "Creating havic: Heterogeneous audio visual internet collection," in *Proc. LREC*. Istanbul, Turkey: ELRA, May 2012.

[27] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[28] D. Vijayasenan and F. Valente, "Diartk: An open source toolkit for research in multistream speaker diarization and its application to meetings recordings," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[29] D. Vijayasenan, F. Valente, and H. Bourlard, "Agglomerative information bottleneck for speaker diarization of meetings data," in *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*. IEEE, 2007, pp. 250–255.

[30] N. Ryant, "Ldc sad," https://github.com/Linguistic-Data-Consortium.

[31] ——, "Diarization evaluation," https://github.com/nryant/dscore.

[32] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[33] "The first DIHARD speech diarization challenge," https://coml.lscp.ens.fr/dihard/index.html.

[34] A. Cristia, S. Ganesh, M. Casillas, and S. Ganapathy, "Talker diarization in the wild: The case of child-centered daylong audiorecordings," in *Proceedings of Interspeech*, 2018.