Longform recordings : Opportunities and challenges

Lucas Gautheron, Marvin Lavechin, Rachid Riad, Camila Scaff, Alejandrina Cristia

Enregistrements de longue durée: Opportunités et défis

Lucas Gautheron[1] Marvin Lavechin[1, 2] Rachid Riad[1, 2, 3] Camila Scaff [4,1]
Alejandrina Cristia[1]
(1) Laboratoire de Sciences Cognitives et de Psycholinguistique, Département d'Etudes cognitives, ENS, EHESS, CNRS, PSL University, Paris, France
(2) CoML Team, INRIA, Paris, France
(3) Laboratoire de Neuropsychologie Interventionnelle, Département d'Etudes cognitives, ENS, INSERM, UPEC, PSL University, Paris, France
(4) University of Zurich, Zurich, Switzerland
alecristia@gmail.com*

RÉSUMÉ

Bénéficiant d'améliorations technologiques récentes, des appareils audio légers et portables sont désormais capables d'enregistrer des dizaines d'heures sans interruption. Nous proposons une description générale de cette technique pour l'étude de la parole, et faisons le point sur ses avantages et inconvénients. Grâce à elle, les linguistes de terrain bénéficient d'un accès unique au langage dans un contexte plus naturel. Cependant, ces enregistrements restent difficiles à annoter manuellement ou automatiquement, en raison de leur durée, du bruit, et de la sensibilité des informations qu'ils peuvent contenir. Des outils *open-source* plus facilement appopriables, auxquels les spécialistes des technologies de la parole peuvent contribuer, favorisent la reproductibilité des travaux des chercheurs. En outre, de nouvelles approches aux techniques d'annotation manuelles ou automatiques rendent cette technique opérationnelle et prometteuse.

ABSTRACT

Longform recordings : Opportunities and challenges

Technological developments have allowed the development of lightweight, wearable recorders that collect audio (including speech) lasting up to a whole day. We provide a general description of the technique and lay out the advantages and drawbacks when using this methodology. Field linguists may gain a uniquely naturalistic viewpoint of language use as people go about their everyday activities. However, due to their duration, noisiness, and likelihood of containing sensitive information, longform recordings remain difficult to annotate manually. Open-source tools improve reproducibility and ease-of-use for researchers, to which end speech technologists can contribute. Additionally, new approaches to human and automated annotation make the study of speech in longform recordings increasingly feasible and promising.

MOTS-CLÉS : enregistrements longs ; validité écologique ; traitement automatique de la parole.

KEYWORDS: daylong recordings ; ecological validity ; automatic speech processing.

Recent years have seen the rise of data collection through wearable, light-weight and unobtrusive devices that collect audio for tens of hours at a time, allowing a uniquely naturalistic viewpoint of language use as people go about their everyday activities. Over nearly a decade, our team gained first-hand experience with the incredible benefits as well as the painful points of this data collection

∗. We acknowledge ANR-16-DATA-0004 ACLEW, ANR-17-EURE-0017; and the J. S. McDonnell Foundation

(a)

(b)

(c)

FIGURE 1 – Examples of wearable recorders. (a) Tsimané child wearing a LENA device in the front pocket of a purpose-made vest. (b) Smart watch recording audio, heart rate, and movement, adapted from Fig 1 in (Liaqat et al., 2018). (c) Body camera on a South Carolina police officer (Ryan Johnson, CC BY-SA 2.0).

technique. By now, our lab has over 20,000 hours of audio, capturing language experiences of over 1,000 children, learning one or more of 16 typologically diverse languages. We provide a brief introduction to this technique, in the hope of allowing our colleagues to decide when it may be a useful tool to add to their kit (for detailed information, see (Casillas et al., 2019)).

1 Interest of the method

1.1 Providing decisive evidence on long-standing debates

Short recordings and controlled data elicitation provide crucial information about language perception and production, but we still know little about spontaneous language use in naturalistic environments. A new window on this was opened by daylong recordings. The technique has already proven fruitful in the field of language acquisition, from where we provide several examples.

One of the key theoretical questions in the field was whether language development is mainly driven by infants wish to communicate, or other processes. Kim Oller and colleagues have been studying development of speech in infancy for many decades, and had discovered that there are speech-like sounds, called protophones, even in young infants – but this still did not settle the question of *why* infants vocalize like this. Only recently, using infant-centered daylong recordings, Oller and colleagues were able to show that, in 6-month-olds, vocalizations were more advanced when the

infant was *not* being talked to, suggesting these vocalizations are endogeneous rather socially driven (Lee et al., 2018). Other results also contradicted prior beliefs that cries were more abundant than protophones at early stages, as the opposite was true even among preterm and fullterm newborns, suggesting in a way that infants are literally born to produce speech (Oller et al., 2019).

The interest for this technique is increasing beyond the language acquisition community. Additional applications being explored include the relationship between social interaction and well-being (Sun et al., 2019), activities among adults suffering from pulmonary diseases (Wu et al., 2018), and measurement of speech and language correlates of neurodegenerative diseases (Riad et al., 2020).

1.2 Drawbacks and challenges

The technique of long-form recordings also comes with its own challenges and limitations. Extracting information from the data may be challenging. The technique produces large amounts of audio which cannot be manually annotated as a whole. Automated tools are thus often required to extract sections and/or annotate the data automatically. However, there are no off-the-shelf solutions for automated annotation, which instead require active development by experts in speech and language technology. Still, even when these tools are developed, they do not have perfect accuracy, and thus it may be impossible to detect small effects. Indeed, contrary to lab-controlled experiments, the data suffer from significant background noise, and are potentially subject to a variety of soundscapes.

Furthermore, recordings might contain confidential or sensitive information, including from people who are accidentally recorded and have therefore not provided informed consent. As a result, researchers often need to ponder difficult ethical and legal questions, for which they may need advice from experts in law and ethics, who may not be familiar with long-form recordings from wearables. Storage on embedded devices or on untrustworthy third-party services such as cloud platforms might require encryption. Data transfers should be secure to prevent leaks beyond the research community.

1.3 When should LIFT members consider using and/or contributing to this technique ?

As with any (new) technique, one should make sure it is appropriate for its research purposes before engaging in it. For readers who are considering **collecting data** with long-form recordings from wearable, we clarify that these recordings are most valuable when (1) ecological validity is key (where e.g. elicitation is inappropriate), (2) unbiased sampling is important, and (3) the phenomenon studied occurs frequently in language use and it is robust to the presence of ambient noise (particularly if automated annotation will be used). We provide some information specifically for LIFT members here (see Casillas et al., 2019 for more examples).

We expect that field linguists in LIFT will find it particularly useful to collect data with wearables when interested in language use in situations where their presence as an observer may not allow a behavior to develop naturally, and when their informants find it hard to report on the use of a form (or their reports may not reflect actual use). To give a specific example, one of our collaborators uses the samples to study patterns of language switching in a highly multilingual community. After her informant has worn the recording device for a whole day, she extracts sections with speech randomly throughout the day, and uses them as prompts to discuss with her informants which language was

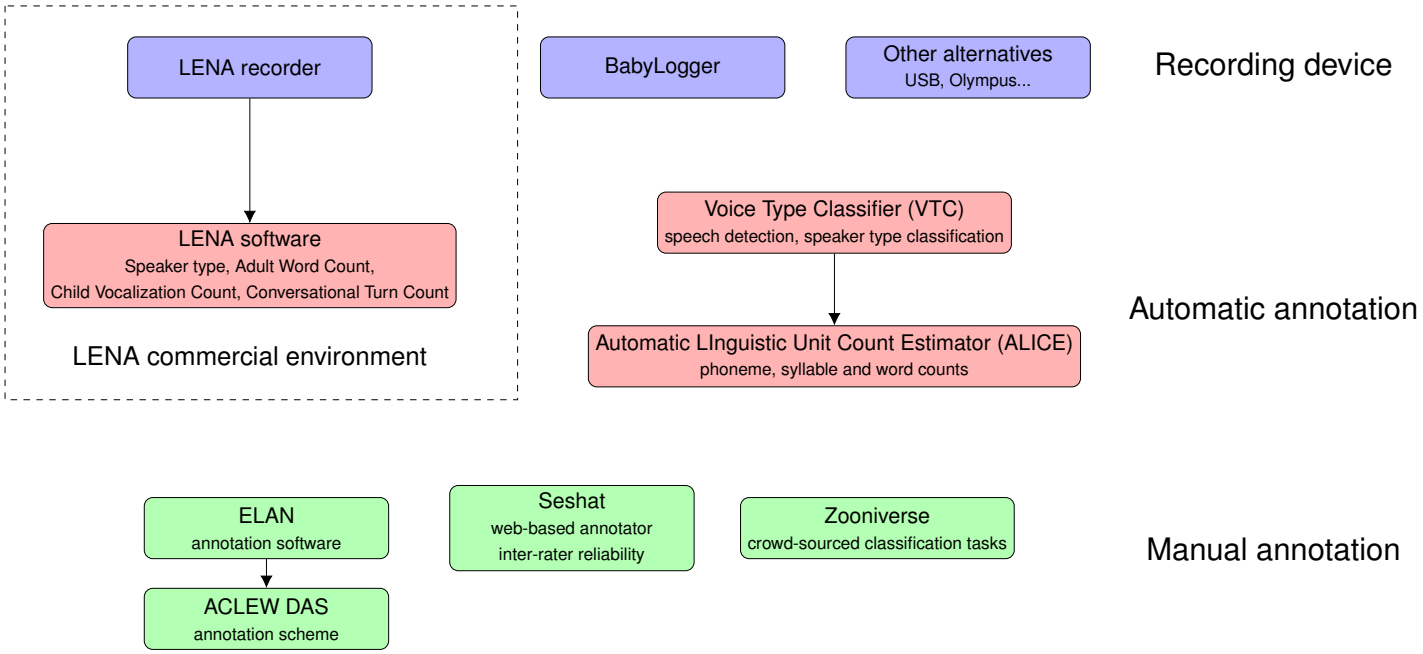

FIGURE 2 – Overview of current solutions for longform recordings, mainly applied to early language acquisition research.

being used, by whom, and why they think that language was used rather than the others (e.g., because of who was overhearing the conversation, or because of the topic).

As for **contributing to the technique's development**, computational linguists in LIFT will find it easiest to contribute to this literature if they already have some experience working with speech, as it has so far proved difficult to use automatic speech recognition (ASR) to generate automated transcriptions. Even a recent study on typical English-speaking adults had humans transcribe the clips, rather than using ASR. But we hope this will not discourage LIFT computational linguists who are interested in this emergent field, as there are many opportunities to start dipping one's toes in, for instance via participation in public challenges (Ryant et al., 2019; Schuller et al., 2019).

2 Tools and ecosystem

2.1 LENA

The most commonly used hardware and software for automatically analysing infant's speech is the Language ENvironment Analysis (LENA©), a commercial lightweight recorder worn in a specially designed vest associated with a closed-source speech processing algorithm. A key strength of LENA is that it provides users with an end-to-end pipeline to collect and analyze daylong recordings, making it efficient and easy to use. LENA's recorders have been designed to be unobtrusive and easy-to-wear to improve the ecological validity of field observations by reducing observer bias. The hardware can record up to 16 hours and can only be analysed by the associated software. Their software was trained on American English input to children aged 0-4 years. It generates an automatic analysis of three key estimates of the child's environment : the number of words spoken by a nearby adult ; the number of times the child made any kind of linguistically relevant vocalization (i.e., speech or babble and

excluding vegetative noises, laughter or crying) ; and the number of exchanges between an adult and the child within a five seconds window, considered as "turns". The LENA technology has been used in numerous studies (Dykstra Steinbrenner et al., 2012; Vandam et al., 2015; Ferjan Ramírez et al., 2020).

The LENA system has been found to be fairly accurate in quantifying children's language environment. A recent study comparing the algorithm automated measures to manual human transcriptions found high correlations for the two measures quantifying the number of adult and child's vocalizations and a moderate correlation for the number of turns (Cristia et al., 2020). Studies on annotations automatically extracted by the LENA speech processing pipeline did not reveal great differences in adult word count accuracy between American English and other languages including Swedish (Schwarz et al., 2017) and French (Canault et al., 2015).

However, there have also been some reports of a problematic level of performance for some of LENA's outputs (notably the key child's vocalization recall) (Cristia et al., 2019). This means researchers using LENA output are relying on a noisy analysis, which potentially hides small statistical effects.

Moreover, despite its use in many linguistic studies and its wide acceptation in the child language community, LENA imposes several limiting factors to scientific progress. There is currently no way to build upon LENA speech processing models as their software is closed source, and gathering information about design choices and their potential impact on performance remains a tedious task. Concerning the recorder, LENA designers' hardware choices cannot be revised. This has raised multiple questions that remain to be answered, mainly : Does a single-channel microphone allow us to capture the full complexity of the child's language environment ? Can alternative models provide us with more faithful metrics of this environment ? Addressing these questions must start by creating open-source alternatives to LENA.

	Autonomy		**Audio properties**				
Device	Battery	Storage	Channels	Sampling rate	Bit depth	Weight	Cost (US$)
LENA	30 h	15 h	1	16 kHz[1]	16	200 g	300
BabyLogger	24 h	SD[2]	4	16 kHz	16	200 g	500
USB	15 h	150 h	1	16 kHz	8	50 g	20
Olympus	25 h	SD[2]	1	22 kHz[3]	32[3]	400 g	300

TABLE 1 – Technical characteristics of various recording devices suitable for child-centered audio collection.

2.2 Alternative tools

We have been leading an effort to build open-source alternatives to LENA speech processing algorithms, providing researchers with models that have similar outputs to the ones returned by LENA, as well as undertaking systematic comparisons of these models with their LENA counterpart. We released a voice type classifier (Lavechin et al., 2020) classifying audio segments into vocalizations

1. Audio undergoes a 10 kHz low-pass filter.
2. Limited by the mini SD card the user fits in.
3. Can be adjusted by the user.

produced by the child wearing the recording device, vocalizations produced by other children, adult male speech, and adult female speech. Building upon this effort, a linguistic unit count estimator (Räsänen et al., 2020) has been developed, allowing users to count the number of words, syllables or phonemes produced by adult speakers. These two models have been shown to outperform their LENA counterpart. We redirect our readers to those papers for more in-depth analysis.

As for the hardware, there exist multiple lightweight recording devices available in the market that one might use to acquire speech, from body mounted cameras to digital voice recorders, each with their own hardware specification (see Fig. 1). There are fewer alternatives to specifically acquire child's speech (Table 1) as these devices require particular safety norms and design to be wearable by young children. One interesting alternative that has been specifically designed for child data acquisition is the BabyLogger (Cao et al., 2018), using an array of four microphones, as opposed to one for LENA. The BabyLogger also performs on-the-fly encryption, protecting the privacy of the participants in case the device is lost or stolen. In the context of patient monitoring, smartwatches (1b) paired with smartphones have also been employed, allowing teletransmission of the data to a remote server at the expense of lesser audio quality (because of bandwidth limitations) and lower duty cycles (to avoid premature battery shortage) (Liaqat et al., 2018). However, more work is needed to know whether or not different hardware specifications might lead to different views on language environments.

2.3 Manual annotations

Because of the noisy nature of the recordings, today's classification algorithms might perform too poorly for practical analyses. Even low-level tasks such as speech detection or diarization can be hard to achieve automatically (Ryant et al., 2019).

Therefore, to evaluate these algorithms for specific corpora and for improving these machine learning models, additional manual annotations are required. In these difficult audio data, human annotations take about 40 times the audio duration. Typical datasets of daylong child recordings can contain thousands of hours of audio, and would require hundreds of thousands of work time to be fully annotated. Nonetheless, it is possible to reduce the amount of audio to annotate manually in a few ways, including by performing random sampling of the data with uniform or non-uniform priors.

We developed a manual annotation scheme to help researchers annotating daylong recordings in a systematic way, thus improving reproducibility and comparisons across studies (Casillas et al., 2017). In a nutshell, our annotation scheme allows researchers to both contribute to machine learning efforts and serve their research goals : Talkers are segmented, and certain layers of information can be added optionally, including transcription or classification into fixed classes (e.g., vocalization type : crying, laughing, canonical, non-canonical).

When adding layers of annotations on top of audio data, researchers face many challenges to handle their campaign of annotations : the *problem around files management* (ex : character-encoding problems, incorrect naming of files), the *non-conformity of the annotations* to the schema established by researchers (misuse of symbols), and the *inconsistency of the annotations* (not properly annotated). That is why we introduced the Seshat software (Titeux et al., 2020). It allows researchers to easily customise and standardize annotations and manage annotators. Finally, to measure how "reliable" are the annotations, we implemented an open-source version of the Gamma Agreement measure in Python (Titeux and Riad, 2020). This allows to measure inter and intra annotator agreement for the type of annotations around speech data.

It would be impractical to rely solely on experts to manually annotate such volumes of audio. Most recently, our team launched a crowd-sourcing project on Zooniverse asking citizens' help to solve simple classification tasks on short audio chunks drawn from the daylong recordings, which proved quite accurate and will allow data annotation at a much larger scale (Semenzin et al., 2020).

3 Conclusion

Despite the many challenges that data from wearables bring, we believe this is a technique fitting to the 21st century, and merits our colleagues' attention as a potential tool in their kit. We highly recommend it to those who are particularly concerned by ecological validity of their observations, and who are interested in phenomena that is common and can be studied from surface (acoustic) features. This is a field in expansion, with at least one speech technology challenge on average over the last 3 years, which is ideal for promoting interactions between speech technologists and field linguists.

Références

Canault, M., Normand, M.-T. L., Foudil, S., Loundon, N., and Thai-Van, H. (2015). Reliability of the language ENvironment analysis system (LENA™) in european french. *Behavior Research Methods*, 48(3) :1109–1124.

Cao, X.-N., Dakhlia, C., Del Carmen, P., Jaouani, M.-A., Ould-Arbi, M., and Dupoux, E. (2018). Baby Cloud, a technological platform for parents and researchers. In *LREC 2018 - 11th edition of the Language Resources and Evaluation Conference*, Proceedings of LREC 2018, Miyazaki, Japan.

Casillas, M., Bergelson, E., Warlaumont, A. S., Cristia, A., Soderstrom, M., VanDam, M., and Sloetjes, H. (2017). A new workflow for semi-automatized annotations : Tests with long-form naturalistic recordings of childrens language environments. In *Proc. Interspeech 2017*, pages 2098–2102.

Casillas, M., Cristia, A., Zwaan, R., and Dingemanse, M. (2019). A step-by-step guide to collecting and analyzing long-format speech environment (LFSE) recordings. *Collabra : Psychology*, 5(1). 24.

Cristia, A., Bulgarelli, F., and 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., Lavechin, M., Scaff, C., Soderstrom, M., Rowland, C., Räsänen, O., Bunce, J., and Bergelson, E. (2019). A thorough evaluation of the language environment analysis (lena) system. *Behavior Research Methods*.

Dykstra Steinbrenner, J., Sabatos-DeVito, M., Irvin, D., Boyd, B., Hume, K., and Odom, S. (2012). Using the language environment analysis (lena) system in preschool classrooms with children with autism spectrum disorders. *Autism : the international journal of research and practice*, 17.

Ferjan Ramírez, N., Lytle, S. R., and Kuhl, P. K. (2020). Parent coaching increases conversational turns and advances infant language development. *Proceedings of the National Academy of Sciences*, 117(7).

Lavechin, M., Bousbib, R., Bredin, H., Dupoux, E., and Cristia, A. (2020). An open-source voice type classifier for child-centered daylong recordings. *Interspeech*.

Lee, C.-C., Jhang, Y., Relyea, G., Chen, L.-m., and Oller, D. K. (2018). Babbling development as seen in canonical babbling ratios : A naturalistic evaluation of all-day recordings. *Infant Behavior and Development*, 50 :140–153.

Liaqat, D., Wu, R., Gershon, A., Alshaer, H., Rudzicz, F., and de Lara, E. (2018). Challenges with real-world smartwatch based audio monitoring. In *Proceedings of the 4th ACM Workshop on Wearable Systems and Applications*, WearSys '18, page 54–59, New York, NY, USA. Association for Computing Machinery.

Oller, D. K., Caskey, M., Yoo, H., Bene, E. R., Jhang, Y., Lee, C.-C., Bowman, D. D., Long, H. L., Buder, E. H., and Vohr, B. (2019). Preterm and full term infant vocalization and the origin of language. *Scientific Reports*, 9(1) :14734.

Räsänen, O., Seshadri, S., Lavechin, M., Cristia, A., and Casillas, M. (2020). Alice : An open-source tool for automatic measurement of phoneme, syllable, and word counts from child-centered daylong recordings. *Behavior Research Methods*, pages 1–18.

Riad, R., Titeux, H., Lemoine, L., Montillot, J., Bagnou, J. H., Cao, X. N., Dupoux, E., and Bachoud-Lévi, A.-C. (2020). Vocal markers from sustained phonation in huntington's disease. *Interspeech*.

Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, S., and Liberman, M. (2019). The second dihard diarization challenge : Dataset, task, and baselines. *arXiv preprint arXiv :1906.07839*.

Schuller, B. W., Batliner, A., Bergler, C., Pokorny, F. B., Krajewski, J., Cychosz, M., Vollmann, R., Roelen, S.-D., Schnieder, S., Bergelson, E., et al. (2019). The INTERSPEECH 2019 Computational Paralinguistics Challenge : Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity. In *Interspeech*, pages 2378–2382.

Schwarz, I.-C., Botros, N., Lord, A., Marcusson, A., Tidelius, H., and Marklund, E. (2017). The LENA system applied to swedish : Reliability of the adult word count estimate. In *Interspeech 2017*. ISCA.

Semenzin, C., Hamrick, L., Seidl, A., Kelleher, B., and Cristia, A. (2020). Towards large-scale data annotation of audio from wearables : Validating zooniverse annotations of infant vocalization types. (Accessed on 11/25/2020).

Sun, J., Harris, K., and Vazire, S. (2019). Is well-being associated with the quantity and quality of social interactions ? *Journal of Personality and Social Psychology*.

Titeux, H. and Riad, R. (2020). *pygamma-agreement : Gamma γ measure for inter/intra-annotator agreement in Python*.

Titeux, H., Riad, R., Cao, X.-N., Hamilakis, N., Madden, K., Cristia, A., Bachoud-Lévi, A.-C., and Dupoux, E. (2020). Seshat : A tool for managing and verifying annotation campaigns of audio data. In *LREC 2020 - 12th Language Resources and Evaluation Conference*, Marseille, France.

Vandam, M., Oller, D. K., Ambrose, S., Gray, S., Richards, J., Gilkerson, J., Silbert, N., and Moeller, M. (2015). Automated vocal analysis of children with hearing loss and their typical and atypical peers. *Ear and hearing*, 36.

Wu, R., Liaqat, D., de Lara, E., Son, T., Rudzicz, F., Alshaer, H., Abed-Esfahani, P., and Gershon, A. S. (2018). Feasibility of using a smartwatch to intensively monitor patients with chronic obstructive pulmonary disease : Prospective cohort study. *JMIR mHealth and uHealth*, 6(6) :e10046.