INTERSPEECH 2017: Show & Tell Contribution
August 20–24, 2017, Stockholm, Sweden

HomeBank: A repository for long-form real-world audio recordings of children

Anne S. Warlaumont[1], *Mark VanDam*[2], *Elika Bergelson*[3], *and Alejandrina Cristia*[4]

[1]Cognitive and Information Sciences, University of California, Merced
[2]Washington State University and Hearing Oral Program of Excellence (HOPE) of Spokane
[3]Duke University, USA
[4]LSCP, Département d'Etudes Cognitives, ENS, EHESS, CNRS, PSL Research University, Paris, France.

awarlaumont2@ucmerced.edu, mark.vanDam@wsu.edu, elika.bergelson@duke.edu, alejandrina.cristia@ens.fr

Abstract

HomeBank is a new component of the TalkBank system, focused on long-form (i.e., multi-hour, typically daylong) real-world recordings of children's language experiences, and it is linked to a GitHub repository in which tools for analyzing those recordings can be shared. HomeBank constitutes not only a rich resource for researchers interested in early language acquisition specifically, but also for those seeking to study spontaneous speech, media exposure, and audio environments more generally. This Show and Tell describes the procedures for accessing and contributing HomeBank data and code. It also overviews the current contents of the repositories, and provides some examples of audio recordings, available transcriptions, and currently available analysis tools.

Index Terms: repository, day-long recordings, language acquisition, child speech, annotation, speech recognition, diarization

1. Introduction

Researchers interested in language acquisition have a long history of documenting language input and output, initially via diaries and eventually through audio or video recordings gathered in the lab or in the home. The last 15 years have seen an emergence of long-form recordings. For instance, the LENA™ Foundation dataset contains 10-16h audio recordings from 329 children aged 0-4 years [1]. Such recordings cannot be shared at present, as they were not gathered with permission for reuse. Fortunately, a number of individual researchers have seen the promise of such an approach and started to collect similar data, this time with data sharing and reuse possibilities in mind. HomeBank houses such data, and tools for their analysis.

2. HomeBank: Goals and benefits

HomeBank [2] (`http://homebank.talkbank.org/`) seeks to promote collaboration across three conceptually separate types of users interested in large scale, spontaneous, and completely naturalistic audio recordings. We will call "Data collectors" those who have privileged access to human participants who agree to be recorded; these are often psychologists, linguists, and cognitive scientists who are well-versed with ethical conduct of research and are able to recruit interested families. We will call "Annotators" researchers who are able to augment recordings with useful annotations, such as segmenting speech from background events and classifying the different speakers on the basis of their individual or class identity. Often, these two roles will be played by the same people, but sometimes annotators may be researchers who do not have the training or access to ethical review boards to collect this kind of data, or they may be researchers who are interested in a population they cannot easily access (e.g., children at high risk of autism). A third group will be called here "Computational or speech scientists", of which there are many subtypes. For instance, some may be interested in extracting statistics from the day-long recordings or associated transcripts to test models of language acquisition; others may use the audio recordings to train or test their speech recognition or other audio classification models. All of them will find themselves creating analysis tools appropriate for the long-form naturalistic recordings.

By using HomeBank, all three types of users can win: They profit from complementarity of expertise, attain greater comparability through use of a common resource, and ultimately gain citations. For example, a speech scientist may apply tools to a corpus collected by several researchers located in different countries, which has been enriched with automatic segmentation and broad annotation, and would have otherwise taken years and several hundreds of thousands of dollars for the speech scientist to collect and annotate. By sharing the tools he/she created with others via the HomeBankCode GitHub portal (`https://github.com/HomeBankCode`), other researchers (perhaps including the original data contributors) can make use of the state-of-the-art speech science measurements. The data collectors and the speech scientist both gain from citation of their respective contributions, by each other and by downstream users.

Moreover, HomeBank, like other data sharing projects (e.g., Databrary [3]), improves over the current *modus operandi*, which is suboptimal at practical and scientific levels. It is currently typical for each research team to collect recordings on their own, develop an in-house post-processing pipeline (including the way of selecting data to transcribe, the implementation of the software for transcription, etc.) and archive their own data and transcriptions, often without much re-use. When such hard-acquired recordings and specialized analysis routines cannot be re-used, the research enterprise as a whole suffers; it increases the cost to society when researchers find themselves reinventing the post-processing wheel. Moreover, such compartmentalized efforts are scientifically suboptimal because researchers base their conclusions on the only samples they have access to, making it virtually impossible to carry out exact replications due to, for example, slight but crucial differences in operationalization of what an utterance is, how vocal maturity is

defined, etc. All of these issues would be alleviated by the use of common data and tool repositories.

3. HomeBank Corpora

Each corpus is a group of recordings donated by a research group, likely collected with a specific purpose. Each contribution consists of two classes of elements, the raw audio recording and an associated time-stamped file with TalkBank-compatible annotations [4], as well as critical metadata (e.g., age of the key child being recorded). Typically, audio data has been segmented and diarized into a few broad classes (e.g., male adult, female adult, key child, other children, overlap, etc.) Corpora are stored under three different sharing levels: completely public, and thus accessible to anyone through a web browser interface (including for download); private and accessible only to HomeBank members, protected via password; or private to a specific subgroup and for specific purposes. Regardless of which section they are in, the use of HomeBank data is governed by the Creative Commons CC BY-NC-SA 3.0 license, and users are required to follow some ethical rules in the use, such as citing the donors, and not distributing the password-protected audio to third parties.

At present, HomeBank contains several corpora, of which one is fully public. Since the Show and Tell is a public event, we cannot show the password-protected corpora there. However, we can say that they contain recordings gathered from infants aged 0-90 months, including some that are at-risk for language delay and other disorders. We can provide more details on the one corpus that has been vetted for public sharing, whose use will be demonstrated at the Show and Tell. The VanDam Public Corpus [5] consists of 159 5-minute files, extracted from day-long recordings collected from 53 target children (37 of whom had documented hearing loss), each of whom wore a recorder in a vest throughout a typical day. These 13.25 total hours have been transcribed orthographically, and they collectively contain 60205 word tokens, and 3951 unique word types.

Researchers can apply for membership using a simple procedure, explained online, and which aims to ensure that users understand the basic principles of ethical use of these potentially sensitive recordings. Attendees of Interspeech will be offered speedy membership.

4. HomeBank's Tools section

As with all TalkBank corpora, HomeBank profits from the power of analytic techniques in CLAN, which are ideal for analyzing text-based transcriptions. The long-form audio recordings also call for other tools. At present we centralize these via a dedicated GitHub group, which contains a number of different repositories. These repositories were developed by individual researchers attempting to attain a number of different goals, and thus employ different programming languages. For instance, Warlaumont contributed a set of Perl scripts that calculates the frequency of key events (speaker segments, conversations) on the basis of the output of the LENA software, and VanDam contributed a MATLAB script that collects acoustic details of selected key segments, such as all female adult vocalizations that are adjacent to a child vocalization (and thus may be child-directed speech).

5. Conclusions

HomeBank is part of a general movement towards more cumulative and transparent research practices. Similar to Databrary [3], we seek to promote the re-use of hard-to-collect developmental data, but our focus differs slightly: Whereas Databrary has specialized in video as a tool to study cognition, our key interest lies in speech and language, and the main kind of data we currently store and analyze are large-scale audio recordings. We also profit from discussions and emergent projects within a network of individuals interested in this kind of data called DARCLE.

One of these emergent projects is the Interspeech 2017 ComParE addressee sub-challenge [6]. Within its general goal of promoting interaction between different researchers interested in naturalistic recordings, HomeBank played an important role by hosting the recordings and annotations used in the challenge. We also expect to be involved in the new TransAtlantic Project ACLEW, in which 9 PIs from 6 countries aim to produce an open-source and augmented analysis pipeline that would do voice activity detection, broad speaker diarization, addressee classification, and other speech analyses. HomeBank will host the audio recordings and may be involved in the permanent storage of analysis code.

HomeBank has come to fill an important need, driven by the emergence of large scale audio recordings gathered in a completely naturalistic fashion. We have specialized in handling the thorny ethical aspects involved in these recordings, while promoting re-use of both recordings and tools. We hope many among the Interspeech community will become HomeBank members to benefit from, and contribute to, these resources.

6. Acknowledgements

HomeBank is funded by a National Science Foundation grant awarded to ASW (1539129), MVD (1539133), and Brian MacWhinney (1539010). We also acknowledge the support of NSF (ASW: NSF-BCS-1529127), Washington Research Foundation (MVD), ANR (AC: ANR-14-CE30-0003 MechELex, ANR-10-IDEX-0001-02 PSL*, ANR-10-LABX-0087 IEC), and NIH (EB: DP5-OD019812).

7. References

[1] J. Gilkerson, K. K. Coulter, and J. A. Richards, "Transcriptional analyses of the LENA Natural Language Corpus," Boulder, CO, Technical Report. [Online]. Available: http://www.lenafoundation.org/wp-content/uploads/2014/10/LTR-06-2_Transcription.pdf

[2] M. VanDam, A. S. Warlaumont, E. Bergelson, A. Cristia, P. De Palma, and B. MacWhinney, "Homebank: An online repository of daylong child-centered audio recordings," *Seminars in Speech and Language*, vol. 37, pp. 128–142, 2016, doi:dx.doi.org/10.1055/s-0036-1580745.

[3] K. E. Adolph, R. O. Gilmore, C. Freeman, P. Sanderson, and D. Millman, "Toward open behavioral science," *Psychological Inquiry*, vol. 23, no. 3, pp. 244–247, 2012, doi:10.1080/1047840X.2012.705133.

[4] B. MacWhinney, "The TalkBank Project," in *Creating and digitizing language corpora*. Springer, 2007, pp. 163–180.

[5] M. VanDam, "VanDam Public Corpus," 2016, doi:10.21415/T5388S.

[6] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom, A. S. Warlaumont, G. Hidalgo, S. Schnieder, C. Heiser, W. Hohenhorst, M. Herzog, M. Schmitt, K. Qian, Y. Zhang, G. Trigeorgis, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2017 computational paralinguistics challenge: Addressee, cold & snoring," in *Proceedings of the Computational Paralinguistics Challenge (ComParE), Interspeech 2017*, in press.