HomeBank: An Online Repository of Daylong Child-Centered Audio Recordings

Mark VanDam, Ph.D.,[1] Anne S. Warlaumont, Ph.D.,[2] Elika Bergelson, Ph.D.,[3] Alejandrina Cristia, Ph.D.,[4] Melanie Soderstrom, Ph.D.,[5] Paul De Palma, Ph.D.,[6] and Brian MacWhinney, Ph.D.[7]

ABSTRACT

HomeBank is introduced here. It is a public, permanent, extensible, online database of daylong audio recorded in naturalistic environments. HomeBank serves two primary purposes. First, it is a repository for raw audio and associated files: one database requires special permissions, and another redacted database allows unrestricted public access. Associated files include metadata such as participant demographics and clinical diagnostics, automated annotations, and human-generated transcriptions and annotations. Many recordings use the child-perspective LENA recorders (LENA Research Foundation, Boulder, Colorado, United States), but various recordings and metadata can be accommodated. The HomeBank database can have both vetted and unvetted recordings, with different levels of accessibility. Additionally, HomeBank is an open repository for processing and analysis tools for HomeBank or similar data sets. HomeBank is flexible for users and contributors, making primary data available to researchers, especially those in child development, linguistics, and audio engineering. HomeBank facilitates researchers' access to large-scale data and tools, linking the acoustic, auditory, and linguistic characteristics of children's environments with a variety of variables including socioeconomic status, family characteristics, language trajectories, and disorders. Automated processing

[1]Department of Speech and Hearing Sciences, Elson S. Floyd College of Medicine, Washington State University, and Spokane Hearing Oral Program of Excellence (HOPE), Spokane, Washington; [2]Cognitive and Information Sciences, University of California, Merced, California; [3]Department of Brain and Cognitive Sciences, University of Rochester, Rochester, New York; [4]Laboratoire de Sciences Cognitives et Psycholinguistique (ENS, EHESS, CNRS), Département d'Etudes Cognitives, Ecole Normale Supérieure, PSL Research University, Paris, France; [5]Department of Psychology, University of Manitoba, Winnipeg, MB, Canada; [6]Department of Computer Science, School of Engineering and Applied Science, Gonzaga University, Spokane, Washington; [7]Department of Psychology, Carnegie Mellon University, Pittsburgh, Pennsylvania.

Address for correspondence: Mark VanDam, Ph.D., Department of Speech and Hearing Sciences, Elson S. Floyd College of Medicine, Washington State University, 412 E. Spokane Falls Boulevard, Spokane, WA 99202 (e-mail: mark.vandam@wsu.edu).

Automating Child Speech, Language and Fluency Analysis; Guest Editor, Brian MacWhinney, Ph.D.

Semin Speech Lang 2016;37:128–142.

DOI: http://dx.doi.org/10.1055/s-0036-1580745.
ISSN 0734-0478.

applied to daylong home audio recordings is now becoming widely used in early intervention initiatives, helping parents to provide richer speech input to at-risk children.

KEYWORDS: Databases, speech production, automatic speech recognition, language acquisition, child language

Learning Outcomes: As a result of this activity, the reader will be able to (1) explain the need for a central repository of daylong family and child audio data and tools and (2) summarize the contributions of the HomeBank database to the scientific and research communities.

To study the speech and language of children and families in a natural context is expensive in terms of the time and effort required to obtain and code raw audio (and video) data. One approach to reduce the cost of data collection is to employ miniaturized, wearable recording devices and subsequent automated speech processing algorithms for processing the raw audio. Recent applications of this approach have led to a proliferation of child and family speech data. However, we have been lacking a central location from which to access and analyze this wealth of new data. The project described in this report addresses this problem directly by offering a central, public repository of daylong family audio recordings collected from the perspective of the child, along with the tools needed to analyze those recordings.

Research on child speech and vocal development solidified in the middle of the 20th century.[1–3] As described by Oller,[4] philosophical changes to the approach of studying child language developed in the 1970s, resulting in widespread attention to child speech and early sound production,[5–7] and the characteristics of maternal speech and mother–infant interactions.[8] The goal of much of this research was to describe and better understand language development starting in infancy. The bulk of the raw data was collected in laboratories or homes using microphones and tape recorders to document specific interactions, short samples, or a few participants. For example, in one early study, continuous 24-hour recordings were collected in the homes of six families with infants 6 to 16 weeks of age.[9,10] The goal of the study was to characterize speech productions of mothers as they interacted with their infants. Manual transcriptions were then made from random segments of the daylong recordings, coding for certain utterance types. The findings of this research explored how mothers used language in their home environment to engage their infants. In another study, Kenyan preschoolers were recorded for 2-hour segments in their home environment using a small body-worn microphone.[11,12] After transcribing the recordings, Harkness found that children who talked more with adults had faster acquisition of language and more linguistic advances.[12] Another study used a radio microphone worn by preschool children to collect speech and environmental audio data over the course of a day.[13] The recordings were collected using a predetermined schedule of 90-second segments at ~20-minute intervals throughout a 9-hour day. The raw recordings were natural, but the coarseness of the automatic recording schedule provided little context for the content of the collected raw audio data. Nevertheless, this work laid a foundation for how linguistic quantity and quality of parent–child exchanges influence educational outcomes. Another well-known study begun in the 1980s recorded 42 families with 1- and 2-year-old children for an hour each month in their homes.[14] The data collection for this study lasted 2.5 years, and it took another 6 years to analyze, code, and transcribe all the material.[15] This work showed that vocabulary growth and word learning are linked to social factors of the families. Specifically, higher family socioeconomic status was associated with larger

vocabularies and improved test performance on standardized language tests, as well as with greater quantities and richness of parental language input to children.

METHODOLOGY OF CHILD LANGUAGE RESEARCH

Child language samples are typically collected either in a laboratory setting or during scheduled visits by researchers to children's homes. In both cases, researchers attempt to elicit productions from children through special games, tasks, and questions. Recordings made in this way are subject to concerns regarding ecological validity and possible biases introduced by the interventions. Furthermore, until recently, the data collected in the laboratory or in the home were subject to severe hardware, software, quality, and storage limitations. Unless substantial effort was put forth, as in the Hart and Risley project lasting nearly a decade, the resultant recordings most often consisted of decontextualized small samples or single-case study designs, limiting the generalizability or extensibility of the findings.[15]

Recent technological advances in computer hardware and software have dramatically changed the landscape of child language development research. It is now possible to collect daylong audio recordings via small, wearable recorders placed directly on participating children. These recordings can be collected in the child's natural environment, providing large, ecologically valid samples of data with consistent formats across laboratories and researchers. Machine learning algorithms designed to operate on these data provide automated detail of certain aspects of the child's productions (e.g., estimates of total syllables produced or total conversational exchanges), the child's linguistic environment (e.g., amount of overlapping talk, gender of adult interlocutors, and so on), and the ambient acoustic environment (duration and amplitude of TV/electronic media, silence, noise). For certain measures, acoustic processing is entirely automated, providing results in a fraction of the time required for a human to transcribe such data. For example, a daylong audio recording (16 hours) collected passively via a recorder tucked into a pocket on a preschool child can be uploaded to a computer and analyzed with speech processing software in under 2 hours. The output includes files with the original audio and a time-aligned, tagged annotation indicating the specified output of the algorithm. Although modern automatic annotation methods are not free from their own weaknesses—the segmentation and diarization of speakers is, of course, not error free—they have opened new perspectives on child speech and vocal development, family dynamics, and clinical assessment and intervention.

AUTOMATIC METHODS OF DATA COLLECTION

The most prominent, pioneering system for collecting and analyzing daylong audio of children and families in their natural, usually home, environments is the Language ENvironment Analysis (LENA) system, first developed in the mid-2000s by the LENA Research Foundation (Boulder, Colorado, United States). The LENA system is currently in use by over 200 universities, school districts, hospitals, and other research institutions (http://www.lenafoundation.org/lena-pro/). Another system for gathering daylong home audio recordings is the Human Speechome Project, which started as a case study of a single child.[16,17] To date, the Human Speechome Project has published some findings and transcription tools,[17,18] but the raw data are not available for use by other researchers, in large part due to privacy concerns. In addition to these systems, there are several researchers who have developed their own tools for collecting and processing recordings within their own laboratories or research settings, but many have not been made publically available or have only been partially described in the literature.

Research utilizing automated analysis of daylong audio recordings using the LENA system is advancing our understanding of autism spectrum disorders,[19–22] childhood hearing loss,[23,24] the consequences of premature birth,[25,26] the role of television viewing in development,[27–30] and more. These studies using the LENA system have utilized thousands of hours of audio data.

The technology is also increasingly used in applied settings, for example in the Providence Talks,[31] Project Aspire,[32,33] and the Thirty

Million Words initiatives.[34] In these projects, the LENA system is being deployed to examine the effects of poverty, hearing loss, and rehabilitation efficacy on young children's developing linguistic systems. These research projects and intervention initiatives leverage the strengths of daylong naturalistic recording combined with automatic speech processing to inform questions of interest to a wide range of groups, including medical professionals, early childhood educators, policy makers, and politicians.

The totality of the work completed or in progress, tools produced, and data collected is not known, and there is no central repository in which unpublished work, tools, and data are stored, organized, or made available to researchers and/or the public at large. Such a central database could be useful not only for child language and developmental researchers, but also for those researchers pursuing technical advances in automatic speech processing including engineers, software and hardware developers, statisticians and data analysts, and computational modelers.

DAYLONG RECORDINGS: TYPICAL COLLECTION AND ANNOTATION

Extended or daylong recordings can be collected and processed using a wide variety of tools. As described previously, starting at least as early as the 1970s, daylong audio recordings have been collected for research purposes. The present project has a critical mass of daylong recordings using a particular technology, the LENA system, but a central repository described here, HomeBank, is designed and intended to be compatible with any recording and associated data of extended family audio. Nevertheless, due to the current paucity of alternatives, the remainder of this section will be devoted to describing typical use of the LENA system.

The LENA device contains a single integrated microphone that records unprocessed acoustic data to onboard solid-state memory. The self-contained recorder is $\sim 1 \times 6 \times 8$ cm and weighs less than 80 g.[35] Unlike most wireless recording setups used in laboratory settings, the recorder does not transmit to a receiver; instead it stores the entire recording on the device worn by the child.

Often, the device is worn by the child in an item of clothing. This allows the recorder to be worn throughout the day, regardless of the location of the child—it can even record during car rides, trips to the park, and so on. The LENA foundation markets custom-tailored clothing options such as vests, which have a pocket sewn into the front into which the recorder is snapped into place. These clothing items have the additional benefits that they have been tested to minimize noise from rustling of clothing and to protect the device from spills while being comfortable for the child and easy for parents to use (cf. recent work looking at acoustic response characteristics of the LENA hardware[36]).

In the typical application, a researcher furnishes a recorder to the parent along with instructions to turn on the recorder when their child wakes up in the morning and turn it off when the child goes to sleep at night. Parents are told that they can pause the recorder, should they need to for privacy, and that entire recordings can be discarded upon request.

When the hardware is returned to the researcher with the recording stored onboard, the audio data then needs to be analyzed. Few laboratories are performing exhaustive transcriptions given the length of recordings being gathered, and instead turn to automatized postprocessing. An ideal goal would be to obtain an orthographic transcription from the acoustic signal. However, this approach has not been entirely successful in natural (i.e., ecologically valid) settings with highly variable vocalizations, overlapping conversations, and various environmental sounds. Thus postprocessing currently focuses on a simpler, though still challenging, automatic segmentation and labeling task: that of breaking up the continuous acoustic signal into segments that are labeled by the primary sound source. This can be done using custom-written scripts, but many LENA users opt to employ the LENA software's automatic labeling algorithms. The LENA algorithms return an audio recording broken down into segments that are assigned a segment label. The labels are organized into around a dozen higher-level categories including talker identity (child wearing the recorder, any other child, adult female, adult male, and human speech overlapped with other speech or nonspeech noise), other sound

sources (TV or other electronic sounds, noise), and silence. The result of this processing is output as a computer text file typically ranging from 20,000 to 50,000 segments per daylong recording. Each line of the output file contains the onset and offset times of the segment, the specific label, and other information such as mean amplitude of the segment. The LENA algorithm output does not provide a written transcription of the words on the recording using automatic speech recognition of the sort found in smartphones. Rather, the LENA system exhaustively segments the recording and provides speaker labels and other categorizations for each segment using categories from the algorithm's predetermined list.[37,38]

There are other types of information that may be derived automatically from daylong recordings. For instance, one can draw estimates of key elements within the child and adult vocalizations, such as the number of words spoken by adults and speech-related child vocalizations (e.g., speaking, babble, singing) versus non–speech-related child vocalizations (e.g., crying, laughing, burping), and whether there are sequences of segments where the key child and an adult alternate (which can be automatically labeled as "conversational turns"). Finally, the LENA system in particular draws from a standardized database to provide even more information by comparing the acoustic features of the child's speech against norms from the Natural Language Study.[39] This allows the LENA software to provide users with an automatic vocalization assessment that aims to provide information about the child's developmental level with regard to speech production ability.[40] The automatic vocalization assessment necessarily relies on normed data, but all other data relies on standard speech technology methods. For example, the LENA system uses a combination of Gaussian mixture models and hidden Markov models to obtain the talker labels and the child speech-related versus non–speech-related labels.[41–43] It uses the open source Sphinx software (CMU Sphinx, Carnegie Melon University, Pittsburgh, PA) as inputs in estimating number of adult words and in producing the child vocalization assessment.[39,44,45]

Supplementing the raw audio and the time-aligned record of labeled segments, several research teams have developed additional software tools that process some or all of these outputs (i.e., the audio file or the text file record) for further analysis.[21,46–48] Several of these researchers have also created practical software tools ranging from database and file management to algorithms for acoustic analyses. The tools are undoubtedly useful to the researchers and their teams, but there was, prior to HomeBank, no central repository that would benefit other researchers or those interested in becoming familiar with the technology. A comprehensive database of external software tools would greatly benefit the research community by reducing the cost of entry and increasing the accessibility of and proficiency with the data. This would especially benefit students and those researchers desiring to take advantage of the LENA system but not currently active.

Data on the reliability of such systems is beginning to emerge, and again the LENA system is the best validated at present. The LENA labels and adult word counts compared with human coders has been published for typically developing children learning English,[41,42,48,49] Spanish,[50] Dutch,[51] and French.[52]

THE PROBLEM OF DISTRIBUTED AND SEQUESTERED DATA

Using modern technological resources of automated data collection, and in particular the growing popularity of the LENA as a research tool, research teams around the world are beginning to amass data sets of naturalistic daylong audio that are vastly larger and more representative than what is currently available in either the Child Language Data Exchange System (CHILDES) repository[53] or Databrary.[54] However, because recordings are collected in natural settings and generally involve families engaged in the full range of typical activities and interpersonal interactions, public dissemination of the raw data raises concerns about violating the privacy of participating families. These privacy concerns combined with the extended duration of the audio recordings have made vetting of the recordings and removal of private information a central, but particularly labor-intensive, concern. As a result,

most of the collected data—estimated to be in the tens of thousands of daylong audio recordings—remains sequestered in individual laboratories. Indeed, some researchers may be required (e.g., by their institutional review board [IRB]) to delete the underlying audio data, not having a system at their disposal to properly vet, store, or secure sensitive recordings.

The development of a system for sharing these rich data collections offers a benefit to the fields of child language acquisition and automatic speech processing. Such a database supports the sharing and development of resources for basic research as well as for educational and clinical work. Facilitating the development of a system for sharing this valuable data as well as improvements on and extensions to the existing analysis tools are the key motivations for the HomeBank project.

INTEGRATION OF HOMEBANK INTO TALKBANK

The system described in this report, HomeBank (available at homebank.talkbank.org/), was designed to be a public repository for daylong family audio recordings, associated files and processing output, and software or algorithm tools for processing data. HomeBank is integrated into the existing framework of the TalkBank databases (which include, for example, the CHILDES database) and tools entailed there. HomeBank thus leverages the massive data and advantages of automatic processing described previously with an existing infrastructure and long history of success of the TalkBank project. Before describing the HomeBank project in detail, CHILDES and TalkBank will be briefly introduced.

CHILDES is a Web-based system for sharing and analyzing child language transcript data.[53] The database at childes.talkbank.org includes 50 million words of transcript data, much of it linked on the sentence level to digitized video or audio recordings that can be played back directly over the Web. Over 95% of the data in CHILDES are publicly available for downloading and analysis without a password. CHILDES is one component of the larger TalkBank system (talkbank.org) that includes additional databases for the study of aphasia, traumatic brain injury, second language acquisition, dementia, conversation analysis, meetings, and other language areas.[55]

CHILDES began in 1984 with support from the MacArthur Foundation and has received continuous support from the National Institutes of Health since 1987 and from the National Science Foundation between 1999 and 2004 and between 2015 and 2019. There are currently 1,800 users of CHILDES located in 35 countries. A search at scholar.google.com reveals 5,482 articles in English that have made use of CHILDES data or programs. However, because this inventory does not include research papers published in other major languages, the actual size of the research literature generated by CHILDES is closer to 6,800 publications. Publications based on CHILDES touch on every major issue in child language, from phonology to intellectual development. The data are most heavily used by researchers in linguistics, psychology, computational linguistics, speech and hearing sciences, sociology, education, and modern languages.

CHILDES and the more general TalkBank system of which it is a component have adopted rigorous international standards for data preservation, documentation, and access. In recognition of this, TalkBank has received the Data Seal of Approval, based on accurate adherence to a set of 16 standards regarding corpus documentation (childes.talkbank.org/manuals/), consistent data formatting in a tightly controlled XML schema (talkbank.org/software/xsddoc/), metadata generation in the Open Language Archives Community (OLAC; http://www.language-archives.org/) and Component MetaData Infrastructure (CMDI; clarin.eu) formats, articulation of a full mission statement, IRB protection (talkbank.org/share), data storage, long-term preservation, Open Archives Institute harvesting of CMDI and OLAC metadata, persistent digital object identification through the Handle System (handle.net), backup systems (mirror sites, archiving, git, and so on), statement of codes of conduct (talkbank.org/share/ethics.html), and proper treatment of copyright (CC BY-NC-SA 3.0). TalkBank and CHILDES are also members of the international Common Language Resources and Technology Infrastructure (CLARIN)

consortium of national language data centers (clarin.eu).

In addition to these achievements as a stable center for the sharing of language data, TalkBank has developed standards, programs, and practices that make it ideal as development site for the HomeBank project. CHILDES provides comprehensive software for analysis of phonological, morphological, syntactic, and discourse features, much of it fully automated. These programs include Computerized Language ANalysis (CLAN)[53] for transcript analysis and Phon for phonological analysis,[56] both with linkages to Praat,[57] a free acoustic analysis software, with special tools for corpus analysis.[58] CHILDES has 30 years of experience in setting up alternative levels of data access (talkbank.org/share/irb/options.html) that protect individual privacy in accord with high-level IRB standards.

HOMEBANK

HomeBank (homebank.talkbank.org) was conceived to offer researchers greater access to raw data and tools associated with the rapidly growing amount of daylong audio files being collected by a wide range of researchers. HomeBank is an effort to address the issue of providing a centralized database or repository for daylong audio files, the associated data and metadata with those files, and software tools for researchers. We expect that the database will be useful to a wide range of users, including those interested in language and child development in the social sciences and those interested in automatic speech processing.

HomeBank consists of a recording database and a code repository. The recording database consists of vetted and unvetted daylong audio recordings. In the vetted section, original daylong recordings and their associated metadata were vetted by experienced, trained listeners to ensure that recordings contained no private or personally identifying information. For example, if a parent recited her name and address on a recording, that audio portion is redacted and inaccessible to the user. The vetted database is unrestricted and open to the public for download, playback, or analysis. However, because the vetted database requires a trained person to listen to the audio in its entirety, and requires that individuals on the recording agree to a more open distribution, this is a relatively modest-sized database.

The other part of the recording database is a restricted database requiring special permissions to access. This part of the database contains unvetted audio recordings and their associated metadata, and recordings for which public distribution has not been agreed to by the participants. Because much of the audio has not been vetted and cleaned by trained listeners, this can be a much larger database than the unvetted database. However, because personally identifying details may be contained on the recordings, several tight user restrictions have been implemented to safeguard the participant families. These recordings are stored in a password-protected computer space only available to registered HomeBank members who have agreed to confidentiality in writing and have passed recognized ethical training on dealing with human data. HomeBank is intended to be accessible and open to the research community while balancing the collective obligation to treating participants ethically. Additional details can be viewed on the HomeBank Web site or by direct email inquiry (contact information at homeBank.talkbank.org).

Metadata are an important aspect of the database that greatly benefits from the extant tools developed from the TalkBank project. In addition to the source audio files and results of the LENA speech processing algorithms, data for each recording generally includes child age, child sex, information about whether the child is typically developing or is a member of a specific clinical population (e.g., having hearing impairment, language delay, autism spectrum disorder, and so on), education level of the child's primary caregivers, country the child was recorded in, dominant ambient language, scores on language or other developmental tests or questionnaires, and comments on the recording. The metadata protocol is designed to be flexible and extensible within the database. For example, certain characteristics of children with autism spectrum disorder may warrant several data fields unique to that population, or children with hearing loss may have audiological data such as audiograms or details of the hearing aid itself.

One especially important category of metadata are a substantial and growing body of human-generated codes for the human analysis of transcripts (CHAT) transcriptions from portions of daylong recordings. Although automatic labeling is valuable because it is computed without supervision and can be efficiently applied to the entire data set, it has several limitations. First, the labels are not as accurate as those that can be made by human listeners. Second, the labels only provide very basic information about the events in the recording. Many researchers are interested in not only when individuals are vocalizing but also in the phonetic, prosodic, linguistic, or semantic content of those vocalizations. Many speech researchers are familiar with traditionally transcribed corpora, but have less experience interpreting machine-generated labels. Thus, the transcriptions are an especially valuable aspect of the database specifically for an important expected user base. In addition, human transcriptions are the gold standard by which machine algorithms are both evaluated and trained, and so are essential to the development of new and improved speech recognition technologies in this domain. We expect that because these gold standard, human-generated transcriptions exist, researchers interested in automatic speech processing will be attracted to the database.

The second part of HomeBank is an open source code repository, HomeBankCode, hosted at GitHub (github.com/HomeBank-Code). GitHub provides free storage and version history. It also has the advantage of being extremely widely used across academia, industry, and hobbyists, making it likely that many potential contributors are already familiar with how to use Git and GitHub and increasing the chances that other users will discover the resource. The repository is publicly available and adheres to the open source philosophy. Code, pseudocode, and standalone algorithms are posted by users, to be modified, improved, changed, or used as the basis for new code by other users. For example, users have developed tools to process LENA daylong audio files, including tools for applying acoustic analyses in batch, performing data cleaning, using transcription functions within CLAN, and deidentifying Interpreted Time Segments (ITS) files. These scripts have been shared through HomeBank for public use and extension. This is the primary source of user-created, postprocessing tools for use on the daylong recordings.

Additionally, the HomeBank website also maintains contact information, an overview of the project, links to related resources on the Web, and samples of related documents such as help with construction of IRB and consent forms for prospective researchers and how to organize metadata.

Anyone with access to the World Wide Web has unrestricted access to the vetted public database and HomeBankCode repository via the homebank.talkbank.org Web site. Before users can gain access to the larger protected database, they provide written evidence of ethical training (e.g., Collaborative Institutional Training Initiative (CITI) certificate), and oral and written agreement of data use and confidentiality, obtained through a HomeBank staff person. Upon registration of membership, the member and supervisees, such as students or laboratory staff registered as working under that member, gain access to the database containing more restricted and unvetted files.

Finally, the HomeBank database is committed to participant respect and beneficence.[59] We have constructed consent form templates that allow parents to opt in to sharing their data in HomeBank. Permission to post the recording is requested at the time of each recording, or some researchers may opt to procure retroactive participant consent to post extant recordings in support of HomeBank. For data to be contributed to HomeBank, consent forms should ask parents to indicate that they approve that their child's daylong audio recordings and metadata are included in a shared database and that they are comfortable with the public being able to listen to their recordings. Participants could alternatively indicate that their data are to be made available only in the restricted data set, for download by more thoroughly vetted researchers only. The consent forms can also provide contact information should they decide at a later date that they wish to revoke access to their data.

WHAT HOMEBANK BRINGS TO THE COMMUNITY

HomeBank provides contributors (of recordings, other data, or code) with a way of augmenting their impact in the research community. Following the model of the CHILDES database, all users of data from either the public or the restricted data sets will be provided with the appropriate citations and be required to cite them in any publications that utilize the data sets. CHILDES has placed consistent emphasis on the importance of citing original sources, with excellent results. CHILDES also maintains methods for citing corpora as publications through assignment of ISBN codes and Handle System identifiers; this policy is extended to HomeBank data as well. Code in the HomeBankCode GitHub repositories is licensed according to the contributor's preferences; for example, contributions to date use the GNU General Public License, version 2, requiring that any derivative work also make source code freely and publicly available.

The data in HomeBank provide an important resource to child development researchers. In contrast to in-laboratory experiments or short home recordings, daylong recordings of children and their caregivers in their natural environments provide more holistic information about child development. They provide a window into all types of children's daily experiences, as well as the ability to better estimate the frequency of different types of events and experiences over the course of a child's day. Many researchers, including student researchers, may have questions that can be addressed by human or automated coding of these daylong naturalistic samples, but lack the time, training, or funds to obtain a large number of original LENA recordings. Even if they do have the resources, those resources may not best be spent on replicating data that has already been collected and archived in the HomeBank database.

Furthermore, a very large database may allow researchers to ask questions of the data that would not otherwise be possible in a smaller database. For instance, users may train listeners working under their supervision to find segments of the recording in which events of a particular type (e.g., singing or book reading) are present and manually transcribe those events, or perform acoustic analyses on selected segments or in certain contexts. Alternatively, users may apply their own automated data analysis methods to a large number of the recordings without necessarily listening to the recordings except to establish reliability. Another advantage of daylong recording is that, assuming there is a way to efficiently process the large quantity of data, it is possible to accumulate data on relatively rare events. In all of these cases, the presence of a well-populated HomeBank database ensures that hard-earned daylong recording data, generously contributed by parents and children for the purposes of increasing our understanding of child development, is put to maximum use.

The raw WAV files included in HomeBank also provide an ample, large data set for input to supervised or unsupervised machine learning systems. For example, unsupervised deep learning neural network algorithms are increasingly being recognized as powerful methods for extracting acoustic features for automatic speech analysis.[60] Such approaches greatly benefit from more and more naturalistic data, which is exactly what HomeBank can provide.

FUTURE DIRECTIONS FOR HOMEBANK

HomeBank was launched in 2015 with support from the National Science Foundation through 2019. TalkBank has agreed to partner with and host HomeBank in perpetuity. The key developments required to maximize HomeBank include increasing the size and variety of the database and making researchers aware of the tool.

There are several developments that will make HomeBank even more useful. First, we may add a third section in the recordings database for daylong recordings that were collected under maximally restrictive sharing conditions, whereby no users outside of the initial laboratory where collection occurred are allowed to listen to them. This might happen, for instance, if they involve families who are particularly concerned with privacy or populations at risk. Naturally, such recordings could be of little use unless another feature was added to HomeBank, namely the development of a system where users cannot access underlying raw

data, but can run queries over it using a scripting interface.

Second, the fact of creating a common repository will allow the community to pool information and gain from the common knowledge. For instance, this common resource would allow the development of even more accurate norms in an open-source format. As mentioned previously, only the LENA system provides users with an estimation of the child's production skills, because only they have made the investment of developing norms, and this only for a representative American sample. As the recording repository in HomeBank grows, new norms can be derived not only for American recordings, but also for those in other countries (provided that researchers there contribute to HomeBank).

Similarly, speech technologists would be able to use the HomeBank recordings to improve the current automatic labeling algorithms. The transcriptions within HomeBank data could be used as labels for training supervised learning systems and for evaluation of automatic speech processing systems. Furthermore, the availability of the audio recordings to approved groups of listeners would enable anyone with the human resources necessary to contribute to providing additional labels that could be used for training their own systems and that would ideally be reshared back to the rest of the community using the recordings. These two examples illustrate the virtuous circle that could be established between the child language and the speech technology community, in essence providing the full set of daylong audio recordings for approved users would provide valuable data to speech processing engineers. This would presumably help them generalize their methods to other types of naturalistic child data. The development of better speech processing methods trained on naturalistic speech would in turn benefit the child development community, because better automated analysis tools will enable more efficient and higher quality research and assessment.

EXAMPLES OF PROJECTS FOR WHICH HOMEBANK COULD BE UTILIZED

In this section we briefly describe several projects, both in general terms and specific projects that would benefit from the HomeBank database. These are selected projects intended to demonstrate the broad utility of HomeBank.

Acoustics of Child and Adult Vocalizations

Several HomeBankCode extensions focus on acoustic analyses to get a more detailed understanding of the sounds the children and adults are producing. For example, Oller and colleagues analyzed child vocalizations by first segmenting them on the basis of amplitude contours into "vocal islands,"[20] which roughly corresponded to syllables, and then analyzed those vocal islands according to duration, spectral tilt, and various other acoustic features. Oller and colleagues then subjected those acoustic features to principal components analysis and used the principal components as inputs to classifiers of clinical group membership. They found that the approach could reliably discriminate the recordings of children with autism from those of nonautistic children with developmental delays and from those of typically developing children.

Other research teams have been combining the output of LENA with acoustic analysis tools available in Praat and custom routines in MATLAB, Python, and R to obtain acoustic information (pitch, vowel formants, spectral characteristics, amplitude, and so on) for child, female adult, and male adult vocalization segments.[57] One example of a current direction of this work is characterizing the acoustics of child-directed speech (CDS) and adult-directed speech during the naturalistic interactions represented in the home recordings. Characteristic CDS has increased pitch, extended syllable and word durations, exaggerated prosody, restricted syntax, and greater phonetic variability.[5,8,61,62] Although CDS in general has had steady attention in the literature at least since the early 1970s, CDS is receiving renewed attention recently due in part to the availability of LENA. This in turn has renewed discussions about ecological validity in this domain. For example, some studies have found that fathers and mothers differ in the speech they direct toward their children.[63–67] This possibility has been examined recently using a very large database of LENA recordings, showing that

in daylong, ecologically valid samples, compared with mothers, fathers used fewer pitch fluctuations,[47,68–70] a greater variety of lexical items, and more complex syntactic forms. Another recent study used LENA recordings and LENA-generated annotated output as the input to doing acoustic analysis in Praat.[71] They found similarities in pitch between mothers and their children, associations between temporal contingencies in conversational exchanges between mothers and children, and acoustic convergence of pitch across conversational blocks of mothers and children.

Characterizing Child–Adult Interaction Dynamics

At a higher temporal level of analysis, researchers have developed tools that use the onset and offset times of child and adult vocalization segments as identified by the LENA software to give a richer picture of the overall pattern of when children and adults are vocalizing over the course of the day and how the children's and adult's vocalizations relate to each other temporally.[72,73] This work has, for example, found that adult vocal responses are more likely when child vocalizations are speech related, and that a child is more likely to produce a vocalization that is speech related when the child's own most recent speech-related vocalization received a response.[21] Furthermore, various components of this feedback loop were found to vary for children of different ages and socioeconomic backgrounds as well as for children who are typically developing compared with those with autism spectrum disorder. Combined with computational modeling work,[21] the results provided support for the theory that there is a positive feedback loop between child behavior and reinforcing adult responses that helps support children's speech development, and that differences in the feedback loop can have cascading effects on the child's overall developmental trajectory.[74,75] These results provide examples of how automatic speaker identification within daylong home audio recordings can provide the quantity of time series data needed to detect the presence of a two-part feedback loop. The code for these analyses was provided as supplemental material to the article and is now available via the HomeBankCode repository on GitHub.

Development of New Tools for Interacting with Data

Other tools being developed include those that make human listening tasks more efficient. Researchers have developed tools for automatically extracting audio segments specified in the LENA output files and the associated WAV file, playing all the sounds from a given talker to the user, and allowing the user to provide feedback, for example on whether or not the sound was assigned the correct speaker label by the LENA algorithm.[49] Other research teams have used hybrid machine–human transcription techniques to test the automatic labeling procedures.[43] Some research groups have further developed custom software for audio extraction and playback, and human listener judgment collection, acoustic analysis, database management, and statistical analysis.[23,27,47] A major advantage to this type of specialized playback software is that it enables much more efficient human coding of the data. VanDam and Silbert reported on human judgments of over 90,000 exemplars of speech segments labeled by the automatic methods of the LENA system, with accuracy/validity compared for different label types.[49] They found that the automatic methods perform similarly to other state-of-the-art automatic speech recognition methods, but that performance varies for different labels (adult men are more accurately labeled than adult women, for example). Acoustic analyses showed that temporal and spectral qualities—but not amplitude—interact in complex ways in the automatic label determination process. This data extraction, playback, and acoustic analysis software has been explicitly developed in a format suitable for sharing, extensibility, documentation, and modification.

DISCLOSURES

None of the authors have conflicts of interest.

ACKNOWLEDGMENTS

This work was supported by a collaborative National Science Foundation grant awarded to

A.S.W. (1539129), M.V.D. (1539133), and B. M. (1539010), and by WSU Spokane Seed Grant Program awarded to M.V.D. It was further supported by a National Institutes of Health grant to Bergelson (DP5-OD019812-01).

REFERENCES

1. Lynip AW. The use of magnetic devices in the collection and analysis of the preverbal utterances of an infant. Genet Psychol Monogr 1951;44(2): 221–262
2. Chomsky N. A review of BF Skinner's Verbal Behavior. Language 1959;35(1):26–58
3. Skinner BF. Verbal Behavior. New York, NY: Appleton-Century-Crofts; 1957
4. Oller DK. The Emergence of the Speech Capacity. Mahwah, NJ: Lawrence Erlbaum Associates; 2000
5. Brown R. A First Language: The Early Stages. Cambridge, MA: Harvard University Press; 1973
6. Lenneberg EH, Chomsky N, Marx O. Biological Foundations of Language. New York, NY: Wiley; 1967
7. Oller DK, Wieman LA, Doyle WJ, Ross C. Infant babbling and speech. J Child Lang 1976;3:1–11
8. Snow CE, Ferguson CA. Talking to Children: Language Input and Acquisition. Cambridge, UK: Cambridge University Press; 1977
9. Korman M. Adaptive aspects of maternal vocalizations in differing contexts at ten weeks. First Lang 1984;5:44–45
10. MacWhinney B. The CHILDES Project: The Database. Vol. 2. New York, NY: Psychology Press; 2000
11. Harkness S. Cultural variation in mothers' language. Word 1975;27:495–498
12. Harkness S. Aspects of social environment and first language acquisition in rural Africa. In: Snow CE, Ferguson CA, eds. Talking to Children: Language Input and Acquisition. Cambridge, UK: Cambridge University Press; 1977:309
13. Wells G. Describing children's linguistic development at home and at school. Br Educ Res J 1979; 5(1):75–98
14. Hart B, Risley TR. Meaningful Differences in the Everyday Experience of Young American Children. Baltimore, MD: Paul H. Brookes Publishing; 1995
15. Hart B, Risley T. The early catastrophe. Am Educ 2003;27(4):6–9
16. Roy BC, Roy D. Fast transcription of unstructured audio recordings. Paper presented at: Proceedings of the 10th Annual Conference of the International Communication Association INTERSPEECH 2009; September 6–10, 2009; Brighton, UK
17. Roy BC, Frank MC, Roy D. Relating activity contexts to early word learning in dense longitudinal data. Paper presented at: Proceedings of the 34th Annual Cognitive Science Conference; August 1–4, 2012; Sapporo, Japan
18. Roy BC, Frank MC, DeCamp P, Miller M, Roy D. Predicting the birth of a spoken word. Proc Natl Acad Sci U S A 2015;112(41):12663–12668
19. Dykstra JR, Sabatos-Devito MG, Irvin DW, Boyd BA, Hume KA, Odom SL. Using the Language Environment Analysis (LENA) system in preschool classrooms with children with autism spectrum disorders. Autism 2013;17(5):582–594
20. Oller DK, Niyogi P, Gray S, et al. Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. Proc Natl Acad Sci U S A 2010;107(30): 13354–13359
21. Warlaumont AS, Richards JA, Gilkerson J, Oller DK. A social feedback loop for speech development and its reduction in autism. Psychol Sci 2014;25(7): 1314–1324
22. Warren SF, Gilkerson J, Richards JA, et al. What automated vocal analysis reveals about the vocal production and language learning environment of young children with autism. J Autism Dev Disord 2010;40(5):555–569
23. VanDam M, Oller DK, Ambrose SE, et al. Automated vocal analysis of children with hearing loss and their typical and atypical peers. Ear Hear 2015; 36(4):e146–e152
24. VanDam M, Moeller MP, Tomblin JB. Analyses of fundamental frequency in infants and preschoolers with hearing loss. Paper presented at: 160th Meeting of the Acoustical Society of America; November 18, 2010; Cancun, Mexico
25. Caskey M, Stephens B, Tucker R, Vohr B. Importance of parent talk on the development of preterm infant vocalizations. Pediatrics 2011;128(5): 910–916
26. Johnson K, Caskey M, Rand K, Tucker R, Vohr B. Gender differences in adult-infant communication in the first months of life. Pediatrics 2014;134(6): e1603–e1610
27. Ambrose SE, VanDam M, Moeller MP. Linguistic input, electronic media, and communication outcomes of toddlers with hearing loss. Ear Hear 2014; 35(2):139–147
28. Aragon M, Yoshinaga-Itano C. Using Language ENvironment Analysis to improve outcomes for children who are deaf or hard of hearing. Semin Speech Lang 2012;33(4):340–353
29. Christakis DA, Gilkerson J, Richards JA, et al. Audible television and decreased adult words, infant vocalizations, and conversational turns: a population-based study. Arch Pediatr Adolesc Med 2009;163(6):554–558

30. Zimmerman FJ, Gilkerson J, Richards JA, et al. Teaching by listening: the importance of adult-child conversations to language development. Pediatrics 2009;124(1):342–349
31. Hodson H. Automatic voice coach gives conversation tips to parents. New Sci 2014;221(2954):22
32. Suskind DL, Graf E, Leffel KR, et al. Project ASPIRE: Spokane language intervention curriculum for parents of low socio-economic status and their deaf and hard-of-hearing children. Otol Neurotol 2016;37(2):e110–e117
33. Sacks C, Shay S, Repplinger L, et al. Pilot testing of a parentdirected intervention (project ASPIRE) for undeserved children who are deaf or hard of hearing. Child Lang Teach Ther 2014;30:91–102
34. Leffel K, Suskind D. Parent-directed approaches to enrich the early language environments of children living in poverty. Semin Speech Lang 2013;34(4): 267–278
35. Ford M, Baer CT, Xu D, Yapanel U, Gray S. The LENA language environment analysis system: audio specifications of the DLP-0121. LENA Foundation Technical Report LTR-03–2. 2008; Available at: http://www.lenafoundation.org/wp-content/uploads/2014/10/LTR-03-2_Audio_Specifications.pdf. Accessed January 25, 2016
36. VanDam M. Acoustic characteristics of the clothes used for a wearable recording device. J Acoust Soc Am 2014;136(4):EL263–EL267
37. Xu D, Yapanel U, Gray S, Baer CT. The LENA Language Environment Analysis System: the interpretive time segments (ITS) file. LENA Foundation Technical Report No. LTR-04–2. 2008; Available at: https://www.lenafoundation.org/wp-content/uploads/2014/10/LTR-04-2_ITS_File.pdf. Accessed March 18, 2016
38. Oller DK. LENA: automated analysis algorithms and segmentation detail: how to interpret and not overinterpret the LENA labelings. Paper presented at: LENA Users Conference; April 2011; Denver, CO
39. Gilkerson J, Richards JA. Impact of adult talk, conversational turns, and TV during the critical 0–4 years of child development. Technical Report LTR-01-2. 2008. Available at: https://www.lenafoundation.org/wp-content/uploads/2014/10/LTR-01-2_PowerOfTalk.pdf. Accessed September 6, 2010
40. Richards JA, Gilkerson J, Paul T, Xu D. The LENA automatic vocalization assessment. Technical Report LTR-08–1. 2008. Available at: http://www.lenafoundation.org/wp-content/uploads/2014/10/LTR-08-1_Automatic_Vocalization_Assessment.pdf. Accessed January 25, 2016
41. Xu D, Yapanel UH, Gray S, Gilkerson J, Richards JA, Hansen JH. Signal processing for young child speech language development. Paper presented at: First Workshop on Child, Computer and Interaction WOCCI; October 23, 2008; Chania, Crete
42. Xu D, Richards JA, Gilkerson J, Yapanel U, Gray S, Hanson J. Automatic childhood autism detection by vocalization decomposition with phone-like units. Paper presented at: Second Workshop on Child, Computer and Interaction WOCCI; November 5, 2009; Cambridge, MA
43. Xu D, Richards JA, Gilkerson J. Automated analysis of child phonetic production using naturalistic recordings. J Speech Lang Hear Res 2014;57(5): 1638–1650
44. Bořil H, Zhang Q, Ziaei A, et al. Automatic assessment of language background in toddlers through phonotactic and pitch pattern modeling of short vocalizations. Paper presented at: Fourth Workshop on Child Computer Interaction WOCCI; Available at: http://www.utd.edu/~hynek/pdfs/WOCCI14.pdf. Accessed January 25, 2016
45. Xu D, Paul TD. System and method for expressive language and developmental disorder assessment. U.S. Patent US8938390B2; January 20, 2015
46. Bořil H, Hansen JH. UT-Scope: towards LVCSR under Lombard effect induced by varying types and levels of noisy background. Paper presented at: Acoustics, Speech and Signal Processing (ICASSP) 2011 IEEE International Conference; May 22, 2011; Prague, Czech Republic
47. VanDam M, De Palma P. Fundamental frequency of child-directed speech using automatic speech recognition. Paper presented at: IEEE Proceedings of the Joint 7th International Conference on Soft Computing and Intelligent Systems and 15th International Symposium on Advanced Intelligent Systems; December 10, 2014; Kitakyushu, Japan
48. Soderstrom M, Wittebolle K. When do caregivers talk? The influences of activity and time of day on caregiver speech and child vocalizations in two childcare environments. PLoS ONE 2013;8(11): e80646
49. VanDam M, Silbert NH. Precision and error of automatic speech recognition. Proceedings of Meetings on Acoustics 2013;19:060006
50. Weisleder A, Fernald A. Talking to children matters: early language experience strengthens processing and builds vocabulary. Psychol Sci 2013;24(11):2143–2152
51. Berends C. The LENA System in Parent-Child Interaction in Dutch Preschool Children with Language Delay [M.A. thesis]. Utrecht, Holland: UMC-Utrecht; 2015
52. Canault M, Le Normand MT, Foudil S, Loundon N, Thai-Van H. Reliability of the Language ENvironment Analysis system (LENA) in European French. Behav Res Methods 2015; 15:1–6

53. MacWhinney B. The CHILDES Project: Tools for Analyzing Talk, 3rd ed. Mahwah, NJ: Lawrence Erlbaum Associates; 2000
54. Adolph KE, Gilmore RO, Freeman C, Sanderson P, Millman D. Toward open behavioral science. Psychol Inq 2012;23(3):244–247
55. MacWhinney B. The TalkBank Project. In: Beal JC, Corrigan KP, Moisl HL, eds. Creating and Digitizing Language Corpora: Synchronic Databases. Vol. 1; Houndmills: Palgrave-Macmillan; 2007:163–180
56. Rose Y, MacWhinney B, Byrne R, et al. Introducing Phon: a software solution for the study of phonological acquisition. In: Proceedings of the 30th Annual Boston University Conference on Language Development. Somerville, MA: Cascadilla Press; 2006:489–500
57. Boersma P, Weenink D. Praat: doing phonetics by computer [computer program]. Available at: http://www.fon.hum.uva.nl/praat/. Accessed January 25, 2015
58. Boersma P. The use of Praat in corpus research. Available at: http://fonsg3.hum.uva.nl/paul/papers/PraatForCorpora2.pdf. Accessed January 25, 2016
59. U.S. Department of Health and Human Services. The Belmont Report: ethical principles and guidelines for the protection of human subjects of research. 1979. Available at: hhs.gov/ohrp/humansubjects/guidance/belmont.html. Accessed January 25, 2015
60. Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. Signal Processing Magazine 2012;29(6): 82–97
61. Fernald A, Taeschner T, Dunn J, Papousek M, de Boysson-Bardies B, Fukui I. A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. J Child Lang 1989;16(3):477–501
62. Hoff-Ginsberg E. Some contributions of mothers' speech to their children's syntactic growth. J Child Lang 1985;12(2):367–385
63. Mannle S, Tomasello M. Fathers, siblings, and the bridge hypothesis. Children's Language 1987; 6:23–42
64. Reese E, Fivush R. Parental styles of talking about the past. Dev Psychol 1993;29(3):596–606
65. Tenenbaum HR, Leaper C. Mothers' and fathers' questions to their child in Mexican-descent families: moderators of cognitive demand during play. Hisp J Behav Sci 1997;19(3):318–332
66. Tenenbaum HR, Leaper C. Gender effects on Mexican-descent parents' questions and scaffolding during toy play: a sequential analysis. First Lang 1998;18(53):129–147
67. Tenenbaum HR, Leaper C. Parent-child conversations about science: the socialization of gender inequities? Dev Psychol 2003;39(1):34–47
68. VanDam M, Strong W, De Palma P. Characteristics of fathers' prosody when talking with young children. Paper presented at: American Speech-Language Hearing Association (ASHA) Convention; November 12, 2015; Denver, CO
69. VanDam M, De Palma P, Strong WE. Fathers' use of fundamental frequency in motherese. Poster presented at: 169th Meeting of the Acoustical Society of America; May 2015; Pittsburgh, PA
70. VanDam M, De Palma P, Strong WE, Kelly E. Child-directed speech of fathers. Poster presented at: Linguistic Society of America 2015 Annual Meeting; January 10, 2015; Portland, OR
71. Ko ES, Seidl A, Cristia A, Reimchen M, Soderstrom M. Entrainment of prosody in the interaction of mothers with their young children. J Child Lang 2016;43(2):284–309
72. Warlaumont AS, Oller DK, Dale R, Richards JA, Gilkerson J, Xu D. Vocal interaction dynamics of children with and without autism. In: Proceedings of the 32nd Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society, 2010:121–126
73. Abney DH, Warlaumont AS, Haussman A, Ross JM, Wallot S. Using nonlinear methods to quantify changes in infant limb movements and vocalizations. Front Psychol 2014;5:771
74. Karmiloff-Smith A. Development itself is the key to understanding developmental disorders. Trends Cogn Sci 1998;2(10):389–398
75. Leezenbaum NB, Campbell SB, Butler D, Iverson JM. Maternal verbal responses to communication of infants at low and heightened risk of autism. Autism 2014;18(6):694–703