DARES-G1: Database of Annotated Real-world Everyday Sounds

M.W.W. van Grootel, T.C. Andringa*, and J.D. Krijnders
University of Groningen, The Netherlands, Email: {m.van.grootel, t.andringa, j.d.krijnders}@ai.rug.nl*

Introduction

The moment the reader reads this sentence he/she might hear the air-conditioning, a chair squeaking, a computer humming and a colleague fiddling. These are examples of everyday sounds and hearing these and interpreting what they caused them is everyday listening: a seemingly simple perceptual ability that has thus-far eluded our technical abilities. Everyday listening can be contrasted to musical listening [1]. Musical listening is studied by psychoacousticians and focuses on the (perceptual) properties of sound such as pitch and loudness that describe the sound, but ignores the events that produced the sound. Compared to musical listening, everyday listening has so far received little attention from the scientific community. One factor underlying this might be the absence of a suitable research database. DARESounds.org, Databases of Annotated Real world Everyday Sounds, is an initiative to collect annotated databases of everyday sounds that are recorded in the full complexity of everyday situations and to make them available to the scientific community. Currently the collection is limited to the DARES-G1 database described in this paper, but more databases are being compiled. DARES-G1, G refers to the city of Groningen (NL), aims to be as good an approximation of everyday sounds as is realistically possible. Additionally, DARES-G1 aims to be suitable for auditory perception research and as validation of (future) everyday sound recognition systems.

Other research communities in auditory/sound research have sound libraries. The Linguistic Data Consortium (`www.ldc.upenn.edu`) for example, is an open consortium of universities, companies and government research laboratories. It creates, collects and distributes hundreds of speech and text databases, lexicons, and other resources for research and development purposes. For research in musical genres and instruments the RWC database [2] is often used. A very interesting research database for everyday sounds exists at the Auditory Perception Lab of Brown University [3]. However, while this database consists of everyday objects undergoing sound producing manipulations, all samples are carefully manipulated, recorded, and selected. This is ideal for analysis purposes, but the samples are not representative for everyday conditions.

There is also an abundance of sound effects and 'ecological sounds' available online (both commercially and free ware such as `www.soundsnap.com` and `www.freesound.org`), but again these are generally unsuitable for (everyday listening) research, due to a lack of annotation and information about the recordings. Because these collections stem from many different origins, the recording protocols and devices will differ and individual sounds may be difficult to compare. In extreme cases this may even lead to characteristic or even signature properties that create the possibility of identifying the recording set-up (a form of musical listening) instead of the content (a form of everyday listening). Finally, many of these sounds are edited and post-processed for aesthetic purposes. This make these sounds less 'realistic', and therefore none of these databases approach the real-life conditions that characterize true everyday listening. For a more comprehensive description of existing data-sources and a guideline to selecting samples see [4].

The challenges of real-world everyday sound recognition

What makes everyday listening so challenging is the lack of any constraints on the input other than that the sound must have been recorded in everyday situations. This is a very weak constraint since most listeners will rarely find themselves in conditions that are not typical of everyday situations. Brushing your shoes on the doormat, doing the dishes, making coffee, and shopping for dinner all correspond to more or less complex patterns of sources (sounds) that are usually easy to recognize if the sources are presented in a normal/realistic context of other sound sources: the sonic environment. However sources may be more difficult to recognize when presented as (short) samples without context. This entails that the source signal itself may not be rich enough for reliable recognition, and that, somehow, the whole acoustic background contributes to target sound recognition. While this may initially appear as a complication, it might be, just as it is for listeners, a benefit since it transfers part of the recognition problem to the domain of reasoning about sounds. This might be of great help when science learns how to utilize this source of information. Annotated databases such as DARES-G1 might be used for this purpose because they provide sources in a rich and original context.

Everyday sound recognition might be a problem that must be solved before a reliable separation between the everyday sonic environment and the speech it contains can be made. One might expect that such a speech detection process is a prerequisite for more detailed speech signal processing and recognition. Databases like DARES-G1 are an essential first step in the development of truly robust speech detection.

*Corresponding author
This research was in part supported by the Dutch Science Foundation NWO under grant 634.000.432 within the ToKeN2000 program

Furthermore, everyday listening constitutes an essential first step towards the eventual assignment of grounded meaning to the speech part of the signal. Because all meaningful speech pertains to possible events in the past, present, or future, the assignment of meaning to speech requires, just as everyday sound recognition, a notion of what is possible in the world and how these possible events are related. While the required knowledge for everyday sound recognition is limited to the relations between sound producing events, preventing meaningless word strings as output from a speech recognition system requires not only language-specific knowledge, but it requires additionally extensive knowledge about "the world", since conversations can refer to all possible events. DARES-G1 therefore also aims at the design of a grounded ontology for sounds to be applied in other recognition systems.

Everyday sound recognition is still a difficult problem. For example the scope of possible sound sources is large, and the set of possible combinations is astronomic. These sounds are "degraded" between source and receiver due to transmission effects such as frequency dependent reflections and damping. This "degradation" represents important information about the acoustic environment which is highly informative (as the acuity of blind listeners proves). Furthermore, just as a context of other sources helps to recognize sources on the basis of insufficient evidence, knowledge about the transmission properties derived from one source, can be used to facilitate the detection and recognition of other sources. DARES-G1 represents this type of information as well.

In spite of the difficulties of the full problem of every day listening, it might be possible to detect subsets of sounds, such as passing cars and the presence or absence of speech with a reasonable performance and without too much difficulty. Systems that recognize verbal aggression in complex social settings are already deployed commercially [5] and constitute a good example of such a subset detector. Systems that detect everyday sounds at a useful performance level can have important commercial applications. The performance of such systems can be assessed using the fully annotated DARES-G1 database.

Requirements

The compilation of database for everyday sounds in real-world conditions involves a number of non-standard considerations.

Uncontrolled recordings – The recordings should typically represent normal day-to-day sounds that can be encountered by anyone. These recordings should be made in a (semi-)uncontrolled way, i.e. neither post-processing, nor influence on the environment to improve the quality of the recording or to remove background sounds are allowed. Events and environments have to be recorded as they are encountered and they should not be engineered other than by selection. This entails that a certain perfect example of an event is unlikely to be part of the database: all recorded events are idiosyncratic and unrepeatable.

The recorded sounds should represent a sizable fraction of the sounds most listeners hear daily. One way to achieve this is to record every n minutes for a number of days while performing normal daily routines. However, this reduces the diversity of the dataset considerably, because the probability is high that a lot of short events, such as brushing one's teeth or taking an elevator will be missed, while (less interesting) longer events, like watching a movie or riding the bus, will dominate. Therefore a selection of 'interesting' events and locations is more efficient to represent the breadth of everyday sounds, at the cost of some ecological validity.

Recording setup – The database must be suitable for automatic sound recognition purposes, but also for perception research. The database must be useful with human subjects to compare the performance of a recognition system with the recognition performance of human listeners, but it must also be suitable to investigate how listeners respond to everyday sounds. The listening experience with headphones benefits from binaural recordings. This type of recording facilitates the localization of sound sources, which improves the ability to hear them separately which in turn helps the identification of the underlying events.

In humans the sound that reaches the ear drums is altered by reflections and diffractions caused by the size and shape of the ears, head and torso. This filtering helps to localize sound sources spatially. The way the sound is altered can be described by the head-related transfer function (HRTF) and differs between individuals. Listening to a sound filtered by a mismatching HRTF will decrease the performance of localizing sounds and is a distraction to the listener, and therefore influences recognition performance in a negative way. When using binaural recording techniques, the goal is to capture this HRTF as well as possible to create the most realistic capture of the world. Unfortunately, the best binaural recording (with microphones near the ear drums) are also the most listener-specific. This makes a standardized database impossible, while a less perfect binaural recording (e.g. keeping only the head-shadow) decreases the ability of localization. However it was found that for intelligibility and recognition, non-individualized HRTFs performed equally well compared to individualized [6]. Therefore, when assembling a database to be used by multiple listeners, a more general binaural recording technique such as on-ear or near-head recordings, is recommended.

Annotations – Detailed annotations are essential; annotations provide a ground truth and should therefore be reliable and informative. The annotation should focus on the source that produced the sound, and should not describe the sound itself. For example, although the word *bang* is a perfectly accurate description for a class of bursts, it is less informative than a description of what caused the sound. So "door closing" or "gun" are better annotations than non-source related descriptors such as "bang".

The use of source related descriptors instead of sound descriptors enlarges the number of classes considerably, because there are many distinct physical events that produce similar sounds. Another issue associated with this form of annotation, is determining which level of description should be chosen. A "bang" might be a single gun shot, but it might also be part of a burst of rapid firing shots, like from an automatic weapon. In that case annotating the entire burst as one event might be more informative than describing each shot separately. Determining the degree of annotation depends on the context of the sound.

Methods

Recording

The semi-binaural recordings are performed with a pair of DSM-1S microphones inside a WHB/N windscreen headband manufactured by Sonic Studios [7]. The setup resembles a small lightweight headphone, which blends well in the environments that are being recorder. The microphones are connected to a Sony MZ-RH10 Hi-MD recorder, which records to uncompressed 16-bit PCM stereo audio files with a sample frequency of 44.1kHz. During outdoor recordings the microphones are head-mounted (by a single person). During indoor recordings the microphones are mounted on a dummy head on a tripod, because the recording author was often involved in the indoor events. The dummy consists of two hardboard oval discs in a cross formation, covered by felted carpet (see Figure 1). The head measures 0.26m in height and 0.17m in width, similar to the head of the author who made the outdoor recordings. When mounted on the tripod, the head stands 1.50m high, while the recording author measures 1.87m high. The recording settings of the MD device are kept constant during the entire procedure at a medium level.

We believe that the use of two different recording setups for in- and outdoor has minimal influence on the quality of the database. First of all, even without considering the recording setup, the recording environments differ vastly on their own, therefore adding another variable will have a small influence on that difference. Secondly, by using an ear-covering set of microphones instead of in-ear microphone, the difference in HRTF will be limited when comparing a human head and the dummy head we created. And, as mentioned earlier, the fact that the focus for the DARES-G1 lies with recognition rather than localization of sounds, the quality of the HRTF is of less importance when listening.

Annotation

Each sound is accompanied by a global description of the content and the location, and timed annotations of the sound sources present in the signal. The annotation is made as soon as possible after the recording, but not *during* because that would introduce noise not specific to that environment. The sound descriptors are used as uniformly as possible without making distinctions between fairly similar sources. For example, a distinction is made between a truck and a car, but not between cars of different manufacturers. Also, sequences of sounds that can be considered to be a single event, are annotated as a group, while distinct events are described separately. Finally sound sources that that are considered a single physical object, but which produce multiple sounds, are described as a whole, unless one of the parts stands out significantly. For example the sound of a bike consists of the sound of the wheels, brakes and gears, but in general are annotated as a single object.

Figure 1: Left: frame of the dummy head used for recording indoors. Right: example set-up for indoor recordings, including the felt-covered dummy, tripod and microphones in the headband.

Group	Locations (number)
Streets	Busy street (9), Quiet street (5), Pedestrian area (2), Bicycle path (3), Residential area (2), Bus stop (3)
Nature	Forest (5), Park (2), Beach (2), Field (1)
Home	Living room (28), Study (13), Flat (13), Kitchen (9), Hallway (5), Bed room (3), Shed (1)
Public buildings	Supermarket (6), Shop (2), Train station (1)
Vehicles	Bus (4), Train (2), Ferry (1)
Other	Lawn (1), Elevator (2)

Table 1: Listing of the available locations grouped by location type.

Description of the database

The database consists of 120 fragments of 60 second recordings, resulting in about 1.3GB of audio. See Table 1 for a summary of the locations that were captured, grouped by general types. In figure 2 an example of

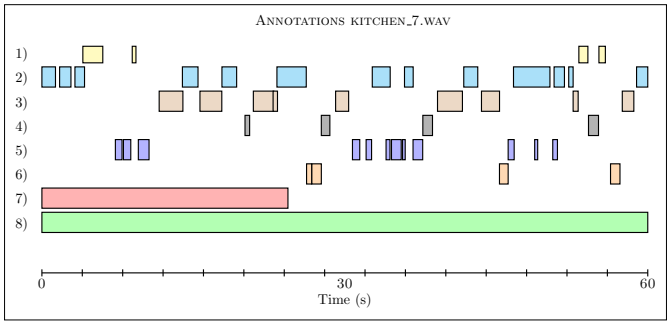

Figure 2: Graphical representation of the annotation of a recording of doing the dishes. Every line represents a different class: 1) *Splashing*, 2) *Scrubbing dishes*, 3) *rummaging through sink*, 4) *Dropping cutlery*, 5) *Clanging cutlery*, 6) *Water drips*, 7) *Fridge*, 8) *Boiler pump*.

the timed annotations and an indication of the of the descriptive detail can be found. The annotation consists of ± 800 different classes, with an average of 25 ± 15 annotations per recording.

All information is stored in an XML file including overall information such as the length of the recording, name of the associated recording, as well as all annotations. The annotations consist of the descriptor of the sound, in English, and the starting and end time.

Discussion

Unlike the currently available databases, DARES-G1 contains the diversity and quality needed for everyday listening research. Enhancements, in the form of additional databases to the DARESounds.org initiative, are more than welcome and can both be aimed at similar recording environments as well as environments that are missed or underrepresented in this collection such as social and industrial areas.

The main problem with creating a database like the one in this paper, is selecting which level of description should be used in the annotations. Being too specific will limit the number of occurrences per sound source, but by using too general descriptions much of the subtleties in the sounds are lost. As mentioned earlier, we solve this by looking at context of the sound and selecting the level of description that seemed appropriate. Another solution is by enhancing the existing annotations with knowledge about the terms used, for example from the lexical database WordNet [8]. That way, both the specific and general aspects can be kept by creating an ontology or semantic network to find related sounds in the database.

Suggested uses

All areas of auditory research that work or want to work with sounds that people encounter on a daily basis, can benefit of the work we presented here. A few possible uses are suggested below. In all cases, proper annotation is a key benefit.

Stress testing and developing recognition systems. Everyday listening system require a large amount of training and test data recorded in diverse, but realistic, environments.

(Computational) auditory scene analysis. Traditionally, and in contrast with what the name of the research field suggests, auditory scene analysis has a focus on constructed and simplified target sounds of speech in noise. Everyday listening databases can extend this to more common sounds.

Auditory cognition research. The rich nature of the database, the realistic quality and the real-world basis, combined with the thorough annotation makes it possible to investigate how people listen to everyday sounds and what the role of knowledge is in the listening process.

Hearing aid validation. Hearing aids are typically testes with pure tones and speech, but not with environmental sounds. Annotated databases with environmental sounds can be used to test the performance of hearing aids on the everyday sounds that the system is most exposed to.

Where to download the database

The DARES-G1 collection is available on the DARESounds website, `http://www.daresounds.org`. The package is 1.3GB and contains the audio files, a Matlab dataset class to facilitate using the data in Matlab and the annotations and descriptions of the files in a single XML file. As more tool and databases become available, they will be distributed from here.

References

[1] William W. Gaver. What in the world do we hear?: An ecological approach to auditory event perception. *Ecological psychology*, 5(1):1–29, 1993.

[2] RWC music database. URL `http://staff.aist.go.jp/m.goto/RWC-MDB/`.

[3] Auditory perception lab at brown university. URL `http://titan.cog.brown.edu:8080/AuditoryLab/`.

[4] Valeriy Shafiro and Brian Gygi. How to select stimuli for environmental sound research and where to find them. *Behavior Research Methods, Instruments, & Computers*, 36(4):590–598, 2004.

[5] P. W. J. van Hengel and T. C. Andringa. Verbal aggression detection in complex social environments. In *Proceedings of IEEE Conference on Advanced Video and Signal Based Surveillance*, 2007.

[6] Rob Drullman and Adelbert W. Bronkhorst. Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation. *The Journal of the Acoustical Society of America*, 107(4):2224–2235, 2000.

[7] Sonic studios. URL `http://www.sonicstudio.com/`.

[8] C. Fellbaum. *Wordnet: An Electronic Lexical Database*. Bradford Books, 1998.