COVER PAGE

Bibliographical reference

The Ambisonic Recordings of Typical Environments (ARTE) Database, Adam Weisser, Jörg M. Buchholz, Chris Oreinos, Javier Badajoz-Davila, James Galloway, Timothy Beechey and Gitte Keidser, Acta Acustica **vol. 105** (Number 4), 2019, pp. 695-713

DOI
https://doi.org/10.3813/AAA.919349

ACTA ACUSTICA UNITED WITH ACUSTICA
Vol. 105 (2019) 695 – 713

Technical and Applied Papers
DOI 10.3813/AAA.919349

The Ambisonic Recordings of Typical Environments (ARTE) Database

Adam Weisser[1,3], Jörg M. Buchholz[1,3], Chris Oreinos[1,2,3], Javier Badajoz-Davila[1,3], James Galloway[2,3], Timothy Beechey[1,2,3], Gitte Keidser[2,3]

[1] Macquarie University, Faculty of Human Sciences, Department of Linguistics, Sydney, Australia.
adam.weisser@hdr.mq.edu.au

[2] National Acoustic Laboratories, 16 University Avenue, New South Wales 2109, Australia

[3] The HEARing Cooperative Research Centre, VIC 3010, Australia

Summary

Everyday listening environments are characterized by far more complex spatial, spectral and temporal sound field distributions than the acoustic stimuli that are typically employed in controlled laboratory settings. As such, the reproduction of acoustic listening environments has become important for several research avenues related to sound perception, such as hearing loss rehabilitation, soundscapes, speech communication, auditory scene analysis, automatic scene classification, and room acoustics. However, the recordings of acoustic environments that are used as test material in these research areas are usually designed specifically for one study, or are provided in custom databases that cannot be universally adapted, beyond their original application. In this work we present the Ambisonic Recordings of Typical Environments (ARTE) database, which addresses several research needs simultaneously: realistic audio recordings that can be reproduced in 3D, 2D, or binaurally, with known acoustic properties, including absolute level and room impulse response. Multichannel higher-order ambisonic recordings of 13 realistic typical environments (e.g., office, cafè, dinner party, train station) were processed, acoustically analyzed, and subjectively evaluated to determine their perceived identity. The recordings are delivered in a generic format that may be reproduced with different hardware setups, and may also be used in binaural, or single-channel setups. Room impulse responses, as well as detailed acoustic analyses, of all environments supplement the recordings. The database is made open to the research community with the explicit intention to expand it in the future and include more scenes.

PACS no. 43.50.Cb, 43.50.Jh, 43.50.Rq, 43.60.Fg, 43.60.Tj, 43.66.Yw, 43.72.Dv

1. Introduction

1.1. Motivation

Over the last two decades, there has been a growing interest in studying human hearing in complex acoustic environments that better represent listening situations experienced in real-life (e.g., [89, 21, 23, 31, 42, 55, 76, 65, 87]). This typically involves using a plurality of sound sources arriving from different distances and directions around the listener, combined with reverberation and ambient noise. Reproducing such environments is particularly informative for the study of hearing devices, since their performance is ultimately assessed by their ability to improve speech perception in noisy real-world environments. However, both speech and noise vary dramatically between the clinic and the real world, to the extent that it is often impossible to predict real-world performance of those devices given the clinical data alone. Moreover, even when data is available about performance in more complex acoustic scenarios, it is often unclear how to generalize these results, and it may be impossible to replicate them. This is because realistic test material in research is collected and reproduced with test-specific requirements, which are usually too narrow to be useful in other fields of research. For example, these may include particular speech-in-noise material, reproduced using an arbitrary loudspeaker arrangement, or designed to test a specific signal processing algorithm (e.g., a directional beamformer of a hearing aid). In addition, these studies often contain technology and materials that are not publicly available.

One way to enable reproducibility and offer tighter experimental control is to use a shared corpus of acoustic scenes, which could be played back in different laboratories on different sound reproduction systems. In this paper, the Ambisonic Recordings of Typical Environments

Received 6 June 2018,
accepted 23 May 2019.

(ARTE) database[1] is described, which is a comprehensive attempt to offer a shared and standardized three-dimensional acoustic scene database. This database was generated using the Higher-Order Ambisonic (HOA) reproduction method, which is a well-studied and theoretically well-grounded method to faithfully reproduce sound fields with low, computable errors [32, 28]. Importantly, HOA audio data may be transformed to different loudspeaker array geometries, as well as to binaural representation. Thus, it should be possible to use these recordings in different laboratories and still retain control over test conditions. This may have the additional benefit of saving the considerable effort required to obtain such recordings, allowing researchers to allocate more resources to their actual research questions.

1.2. Application in Hearing Research

Hearing perception phenomena are traditionally studied in isolation, by carefully controlling acoustic stimuli and varying them along a few dimensions, such as signal-to-noise ratio (SNR), amplitude modulation and reverberation time [61, 60]. In practice, hearing takes place in uncontrolled settings [33], where multiple signals compete for the listener's attention. The listener segregates, integrates, localizes, and understands different targets in auditory scenes with highly variable SNRs [19]. Bridging the knowledge gap between the two extremes – reality and clinic – is not straightforward, but increasingly results in methodologies that try to emulate realistic listening in the laboratory [25, 33, 59]. When it is done well, it has the advantage of achieving increased ecological validity[2] as well as retaining experimental control over the stimuli [65]. This is in comparison with field tests, for example, that are more ecologically valid than clinical setting, but cannot be well-controlled and are not generally reproducible. Ideally, this control should enable the elucidation of the correspondence between the traditional, highly-controlled studies, to the real world. A first step in this direction is a database of complex acoustic environment recordings that has been announced recently by [34]. The database contains multiple everyday scenes that were binaurally recorded by people wearing ear-level microphones. Scenes include restaurants, public transport, and walking in the city, and the recordings were supplemented with derived acoustical data, such as their frequency and modulation spectra, and interaural time and level differences. The intention of this database is to facilitate binaural hearing research, particularly with hearing-impaired listeners. The present ARTE database was designed with a broader aim in mind – to provide an experimental research tool that enables testing of realistic hearing in realistic controlled (clinical) conditions. Instead of binaural recordings, a microphone array was employed in combination with the HOA method to allow the reproduction of real-world scenes using appropriately positioned loudspeaker arrays, as well as binaural playback over headphones. By using loudspeaker-based reproduction, subjects can utilize their individual spatial cues, including head movements, and hearing devices can be more easily integrated.

Utilizing realistic listening environments in hearing research may affect outcome measures in different ways. For instance, the ability to understand speech and to communicate with others will be affected by the noise and reverberation that is introduced by the given environment [54]. In particular, the temporal, spectral, and spatial variability of the noise may change the instantaneous SNR of the speech signal that is received at the two ears [31]. Normal hearing listeners can take advantage of these SNR fluctuations in speech intelligibility either by within-ear glimpsing [24] or better-ear glimpsing [22]. Additional benefit may be provided by the binaural auditory system utilizing interaural time difference cues to spatially unmask the target speech (e.g., [29]). However, the benefit of any of these mechanisms may be reduced by the presence of room reverberation [50]. In addition to these signal energy-related mechanisms, non-target talkers in the environment can also impair intelligibility by distracting the listener, and thereby challenge cognitive mechanisms such as selective attention. Even though these informational masking effects have been widely studied in the laboratory [48, 47], their real- life relevance is unclear (e.g., [85]). Given the complexity of all these environment-specific acoustic factors together with the limited understanding of their combined effect on hearing outcomes, their accurate reproduction in the laboratory is important, in particular for assessing functional hearing abilities in hearing-impaired listeners.

Hearing-impaired listeners are considerably more susceptible than normal- hearing listeners to environmental effects on speech reception (e.g., [40]), which may be either due to their reduced auditory sensitivity, frequency selectivity, or temporal acuity, or due to an age- related reduction in their cognitive function. Either way, the real-life auditory and cognitive mechanisms are currently not well addressed in laboratory-based speech-in-noise tests. In such tests, semi-anechoic target sentences are presented in masking noise that consists of speech- shaped broadband noise (e.g., [67]), speech babble, or speech noise (e.g., [17]), and is presented at a prescribed level from a few loudspeaker directions (e.g., [52, 62, 66]). The listener's score of correctly identified speech is then used to adapt the speech level (i.e., the SNR) until 50% intelligibility is achieved. The resulting speech level, or the speech reception threshold (SRT), is often obtained with ecologically invalid levels of speech and noise, which may be at least partially explained by the rather artificial speech and noise material that is not actually encountered in reality [65, 76]. Hence, providing ecologically valid noise-level scenes may be a necessary stepping stone to direct

[1] The latest version of the database can be found here: https://doi.org/10.5281/zenodo.2261632.

[2] According to Bronfenbrenner [20]: 'Ecological validity refers to the extent to which the environment experienced by the subjects in a scientific investigation has the properties it is supposed or assumed to have by the investigator.'

research efforts to acoustics that do not just stem from a clinical convenience, but are also encountered in reality.

Apart from assessing speech recognition performance in realistic noise, the application of realistic acoustic environments may be important also when sounds other than speech are the main signal of interest (e.g., [49]). The above-mentioned auditory mechanisms are essential in forming a full image of the environment surrounding the listener – e.g., sources behind the head, outside of the visual field – and helps the listener to direct attention to a source of interest (e.g., a talker), or to avoid being hit by an approaching object (e.g., a car). The combined listening experience that occurs in the real-world may be explored using the complex acoustic stimuli provided by the ARTE database.

Differences between typical laboratory and real-world conditions also have consequences for the operation of hearing devices (e.g., [16, 23, 74]). Modern hearing devices are equipped with a host of nonlinear algorithms with the primary aim to increase the effective SNR of speech in noise, through algorithms for noise reduction, microphone directionality, and dynamic range compression [45]. These algorithms are often validated under controlled laboratory conditions, but because of their inherent nonlinearities, they may respond differently to the speech and noise signals experienced in the real life and, as a result, deliver uncertain outcomes [65]. Furthermore, realistic acoustic environments can be influential in terms of subjective pleasantness, comfort, or the lack thereof [13, 88]. A pleasant sounding environment can be conducive for conversations, music listening, and be psychologically comfortable. However, hearing aids can also affect perceived acoustic comfort in a way that is unlikely to be captured in clinical settings [75]. Therefore, the design of future hearing aids should benefit from testing them under more realistic conditions.

1.3. Existing Databases

As motivated and described above, a real-world acoustic environment database that is suitable for the given application in hearing research requires at least that (i) sound files are provided in a format that allows spatial reproduction of the recorded scenes via arbitrary loudspeaker arrays and headphones with an adequate accuracy, (ii) the sound pressure level is provided to allow the reproduction of each recording at its original level, (iii) some basic acoustic and other scene descriptive information is provided that allows informed selection and comparison of scenes, and (iv) RIRs are provided that allow to add speech (or other sound) material to the scenes, as for instance required for implementing a speech-in-noise test.

There are several types of databases that have been presented in literature that are relevant for hearing research and address some of the above requirements, but, to the best knowledge of the authors, none of them satisfies all these requirements. Some databases focus on typical acoustic conditions of specific places and provide detailed descriptions, average values of various acoustic measures, and sometimes room impulse responses. Other databases provide actual sound recordings in different audio formats, but lack important acoustic and other descriptive details. Below is an overview of the existing relevant databases, their intended usage and their limitations for applications in basic and clinical hearing research.

Noise level surveys: Only a handful of studies have been published that surveyed typical noise levels in various daily scenarios. All of these studies sampled the acoustics using ear-level recordings in some manner. They vary in the choice of scenarios, the detail of the derived acoustical data they provide, and the depth of the accompanying acoustical analysis. [72] published the first comprehensive survey of noise and speech level distributions in schools, homes, hospitals, departments stores, trains and airplanes. The A- weighted broadband noise levels and standard deviations were reported, but with no details about other acoustic parameters of the noises and environments. [42, 81, 76] each observed 18- 20 subjects in their daily acoustic environments, who were reportedly satisfied, experienced hearing aid users. [42] investigated the subjects' reaction to their "auditory ecology" and recorded using two omnidirectional microphones, mounted on behind-the-ear dummy hearing aids. The broadband noise levels (in dB) of ten environments are provided along with their standard deviations. The data is also divided into situational categories (conversations, TV, music, etc.), rather than location categories only. Subjects also ranked the relative importance of these daily environments to themselves. [81] conducted a similar survey, but presented a much more detailed statistical and acoustical data set. Analysis of recordings of eleven broad scenario categories included their mean octave-band spectra, group broadband level percentiles, mean and standard deviations, frequency of occurrence and some relevant estimates using subjective data. In [76], a particular emphasis was put on estimating the SNR in various daily situations, divided into nine typical categories. Conveniently, the mean and distribution for the better and worse ears were provided in dB and dBA, as well as the power in octave bands. Similar data was reproduced in [87]. In parallel, surveys in the noise control literature have published levels of a large number of environments and situations. The Noise Navigator™ Database [15] compiled over 1700 levels of different objects and scenarios. Mean or maximum sound pressure levels are provided along with the distance from the source during the measurement. In the more recent Non-Occupational Incidents Situations and Events (NOISE) database the focus was on leisure activities and related statistics about noise exposure levels, mean, maximum levels and standard deviation, along with a description of the measurement conditions are provided [14]. As the primary goal of these studies was to survey daily acoustic environments, they tend to suffer from lack of specific descriptions of the environments that belong to the particular categories. While the derived acoustic data may be sufficient to roughly model an equivalent steady-state reference environment to the categories in the abovementioned studies, there are no

actual recordings available of archetypical scenarios. Furthermore, there are no mentions of the spatial distribution of the various sources, the room acoustic characteristics of these environments, the exact listener positions, or the temporal dynamics of the acoustic scenes.

Soundscapes: In soundscape studies researchers have looked into detailed acoustic and psychological characteristics of particular environments, such as restaurants [51] and public squares [18]. Unfortunately, integrating the output from theses studies into a single comprehensive database of daily scenes may be impossible, because they are not standardized. Moreover, soundscape studies do not always avail the recordings for public use. Nonetheless, a number of publicly available soundscape databases do exist with a focus on the sonic and environmental qualities of the scenes, rather than the acoustical ones. Notably, one such database is the World Soundscape Project Database [80]. It includes comprehensive descriptions of the recorded environments, area photos, exact map locations, and sometimes a timeline description of different sound events throughout the recordings. However, whether single or multichannel, e.g., [7], these recordings are generally uncalibrated, so that they may not be useful for controlled clinical research of the kind that is required in speech communication or hearing device work. Similar environments are sometimes used in the study of environmental sounds (other than speech and music), where the scenes provide the backdrop for the target sound. The Database for Environmental Sound Research and Application (DESRA) was an attempt to provide a systematic source of such sounds [37, 4].

Room impulse response databases: Several audio databases were released for public use that are primarily intended for evaluating different acoustic speech enhancement methods, which mainly focus on the reverberant characteristics of realistic acoustic environments as captured by the room impulse response (RIR). These databases can be used to synthesize realistic scenes by superposing prerecorded anechoic speech and noise signals that are convolved with the provided RIRs, but they are not well suited for reproducing the full complexity and dynamics of the environments experienced in the real world. [46] recorded multichannel head-related and room-related impulse responses using in-ear and behind-the-ear microphones on a manikin and humans [2]. They included an anechoic chamber, an office, a courtyard, and a cafeteria, and provided the reverberation times of these environments. However, the database specifically pertained to the quiet room conditions, rather than to the occupied locations. The Multichannel Acoustics Reverberation Database at York (MARDY) provides RIRs of different configurations of reflectors and absorbers in a variable acoustics rooms [84]. The RIRs were recorded using both an omnidirectional microphone and a eight-element linear microphone array at three different source-receiver distances, and was specifically designed for testing de-reverbration algorithms [8]. Similarly, in the Aachen Impulse Response Database (AIR) [43] binaural room impulse responses (BRIR) are provided for four different rooms in different source-receiver configurations [1]. Single source recordings were made with a dummy head, with the explicit aim to be employed in studying de-reverberation algorithms. Specific descriptions of the room dimensions and reverberation times are provided as well. In a more room-acoustical focused approach, another database provides BRIRs in hundreds of receiver locations within three rooms with fixed source position, in order to have a dense mapping of the source-receiver acoustics [78]. The recordings were done both in omnidirectional and B-format configurations [12].

In other recent databases, the main aim was to support beam-forming algorithms and the absolute acoustic measures were not reported. The Multichannel Impulse Response Database (MIRD) contains impulse responses (IRs) of microphone arrays in a variable reverberation room [39, 9]. Different geometrical configurations of linear arrays of eight microphones were used to obtain the responses of loudspeakers on two semicircles around the array. In another database, the scenario was one specific medium-sized room with four target talkers seated around a table and live babbling talkers (0, 8, 24 or 56) surrounding them [86]. Twenty-four microphones were placed in different locations on and between the talkers and obtained different combinations of talkers and speech-babble, in addition to the head- related impulse responses of the room [3]. Finally, the Open Acoustic Impulse Response (OpenAIR) library offers a platform for sharing the IRs that were recorded in different locations using various methods, including multichannel B-format measurements [64]. The database also includes exact map locations, photos, and room acoustical data derived from the measurements [11].

Machine learning databases: The last class of databases reviewed here serves the design and training of machine learning algorithms that perform scene classification and event identification of audio recordings. For example, the yearly Detection and Classification of Acoustic Scenes and Events challenge (DCASE) [5] utilizes the TUT[4] databases. The 2017 database [56], for instance, contains a closed set of 15 scene classes (e.g., home, park, train, grocery store). Within the challenge, new algorithms were set to compete with a baseline level of 61% successful classification (averaged over all scene classes) of 10 s long segments, which were edited from 3-5 minute binaural recordings of the previous year [58]. Similar databases were released in the past (see [57] for a complete list and a review). Two relevant multichannel databases in this category are the Multi-channel Acoustic Noise Database (DEMAND), which contains acoustically calibrated 16-channel recordings of everyday scenes of six broad classes: domestic, office, transportation, public, nature, and street – each containing three scenes [79, 3]. Finally, the EigenScape is an database of everyday scenes

[3] A 5-min excerpt of that 24-channel database recording is available [10].

[4] Tampere University of Technology

that specifically aims to serve applications of classification, which was inspired by soundscape research requirements, and provides full 3D fourth-order ambisonic recordings in eight classes (eight scenes each): beach, busy street, park, pedestrian zone, quiet street, shopping center, train station, and woodland [35, 6]. Unfortunately, no calibration values or detailed information about the ambisonic processing are provided for EigenScape. Notably, this class of databases contains large amounts of recordings with rich annotation that can be used to robustly train the classification algorithms. However, despite the laborious surveying and annotation done to generate them, they are mostly unsuitable for auditory research because they are uncalibrated in level and other acoustical data are missing.

1.4. Goals

From the above discussion, it is evident that no audio database or acoustic survey exists that allows the faithful reproduction of the acoustic environments experienced in the real world, as required for conducting hearing research with highly ecologically valid outcomes. To at least partially address this limitation, the multichannel acoustic scene database, ARTE, is provided here, which was designed with the following goals:

1. Provide to the research community accessible multichannel recordings of a range of realistic acoustic scenes that can be: a) used in a large variety of auditory perception tests with improved ecological validity; and b) played back in loudspeaker arrays of different geometries as well as under headphones.
2. Enable standardization and replication of auditory perception tests that utilize realistic noisy environments.
3. Complement the multichannel recordings with measured multichannel RIRs as well as basic derived acoustic data.

Goals 1b and 2 are addressed by selecting the HOA format as reproduction method (see Sections 1.1 and 2.4) and Goal 3 is addressed by conducting RIR measurements in all environments (see Section 2.2) and deriving the required acoustical data from the measured RIRs and HOA recordings (Section 3). Goal 1a is (partially) addressed by the perceptual evaluation presented in Section 4, but ultimately, will be revealed when the ARTE database is utilized in future hearing research.

2. Methods: Database Generation

The process of selecting, recording, processing, analyzing, and preparing of the acoustic scenes for file sharing is described below.

2.1. Scene Selection

The recorded environments were selected with the intention to cover a broad range of typical everyday situations that take place in diverse acoustic conditions. Recordings were obtained at a variety of private and public spaces in Greater Sydney, such as cafés, a train station, a library, an office, food courts, a living room, and a dinner party. These locations were considered common enough by the authors, mostly appeared in previous studies that surveyed hearing aid usage [76, 42], but also had the physical and technical conditions for the recording equipment to be set up. As such, the present set of scenes were all situated in urban settings, of Western, English-speaking environments. The recordings captured several hours worth of daily activity in these locations, including incident conversations, footsteps, machinery noise, vehicles, amplified sounds, animal sounds, and others. For the current database, two minute excerpts were selected out of these recordings due to storage space limitations. In each environment, the microphone array (see Section 2.2 below) was centrally positioned in the recorded space, and the microphone look-out direction (0°) was directed to the direction of interest in the scene. For specific details about individual scenes, see Section 3.1.

2.2. Sound Reproduction Format

As HOA technology has matured over the last years, suitable systems for its loudspeaker reproduction are being set up in more and more laboratories around the world. The application of HOA is widely established in auditory displays, virtual reality, and sound engineering, and is increasingly used in hearing research. Using HOA, the sound field is either synthesized in software or recorded using microphone arrays. While the specific system (microphone and loudspeaker arrays suitable for HOA) required for hearing research is not standardized at present, the HOA technology itself offers a high degree of compatibility, which may alleviate the need for hardware standardization. This is because reproduction on any HOA system is based on the same acoustical principles that enable the (re-)synthesis of the intended sound field, within a known error margin (see Section 2.4). Therefore, the standard reference for these systems are the physical sound fields themselves. However, even though the physical sound reproduction error is known or can be measured for any HOA system (e.g., [70]), the accuracy required for hearing research is not clear and may depend on the details of the applied acoustic scenario, as well as on the actual auditory mechanism or hearing ability under assessment.

Having loudspeaker systems that are capable of reproducing three-dimensional sound fields also makes them attractive for hearing device research (e.g., [16, 26, 36, 71, 70]), where wearing headphones is mechanically, acoustically and ecologically too far from realistic aided listening conditions to attest for aided spatial perception of sound. Additionally, methods that aim to perceptually restore the sound sources, but not the physical sound field, may fail to realistically simulate the function of hearing device beamformers, which work differently than the human ear. Because of this applicability for hearing device research, as well as the relative independence on the specific technical setups used (recording and reproduction are separate), the HOA 3D sound field reproduction method was applied in ARTE.

2.3. Recording and collection of material

2.3.1. General scene recording practices

For each recording done in a public location the authors first obtained the necessary clearances from the relevant property managers. The recordings were always attended by at least one person to ensure that no passers-by came too close to the microphone array, or to the rest of the equipment, and a number of clearly visible signs were placed to inform about the recordings. Curious passers-by were told that their conversations may be picked up by the system and played back in research settings. A minimum distance of about 1.3 m between any acoustical source and the microphone array was maintained to avoid spatial distortion in reproduction due to near-field effects.

Several hours worth of recordings were typically obtained in all locations, with the recording time window chosen so that it enabled capturing different levels of activity of the location. Also, it had to include the possibility to perform room impulse response (RIR) measurements in conditions that were as quiet as possible, which usually meant obtaining the permission to stay in the premises before or after business hours. Recordings in outdoor locations were only conducted under fair weather conditions with no wind or rain to avoid damage to the equipment. The only outdoor recordings that were carried out were nevertheless corrupted by wind-noise and had to be excluded from the database.

2.3.2. Multichannel recording equipment

The acoustic scenes contained in the ARTE database were recorded with a microphone array that was designed and built at the National Acoustic Laboratories. The array consists of 62 miniature microphones (Knowles FC-23629-C36) that are flush-mounted inside rubber seals[5] on a hard plastic sphere with a radius of 0.05 m. The microphones are arranged symmetrically in rings, with 24 microphones in the horizontal plane with an angular separation of 15°, ten microphones at both +25° and −25° elevation with an angular separation of 36°, six microphones at both +50° and −50° elevation with an angular separation of 60°, and three microphones at both +75° and −75° elevation with an angular separation of 120°. Inside the plastic sphere, each microphone is connected to a separate preamplifier, which applies a gain of 14 dB and provides balanced outputs. The 62 output channels are then connected to two 32-channel RME M-32 analog-to-digital converters that are linked to a silent measurement computer via an RME MADI HDSPe sound card. The microphone array was calibrated such that the original sound pressure level of the scene was maintained during reproduction. Recordings were done with Audition 3 (Adobe Systems Inc.), at a sampling rate of 44.1 kHz and depth of 32 bits, and all subsequent editing and post-processing was done using MATLAB (The Mathworks, Inc.). This recording fidelity is considered sufficient for the HOA encoded recordings, which exhibit inherent reproduction errors above 7.4 kHz due to spatial aliasing of the 62-element microphone at high frequencies [71].

[5] This was done to reduce vibrational coupling between microphones and chassis and also prevents air leakage around the microphone inlet.

2.3.3. RIR measurements

Multichannel RIRs were measured in all the environments to allow the integration of additional sound sources to the recorded scenes during post-production. This is necessary, for instance, when the recorded scenes are used as background noise in a speech test. In such case, anechoic target speech material can be convolved with the RIRs and added to the recorded scenes at a given speech level or SNR. The resulting target speech then includes the reverberation that is characteristic for the given environment, which is important for the perceptual integration of the target speech and the background noise. A MATLAB routine generated a 20 s long logarithmic sweep signal [63], which was played through a DSE A2760 amplifier that drove a Tannoy V8 loudspeaker. The multichannel RIR was automatically calculated from the recorded sweep response. The resulting RIR was subsequently displayed, to allow manual verification that the dynamic range was at least 50 dB and that the signal decay was clean. Otherwise, where possible, the RIR measurements were repeated until the response appeared clean enough. The absolute sound pressure level of the sweep was around 90 dBSPL at a distance of 1.3 m, but depended on the given location. In most locations the RIRs were measured in quiet without any people around, but in some locations this was not possible. There, the sweep had to be reduced to more comfortable levels and increased ambient noise level of people and traffic was inevitable. In loud measurements, bystanders were advised to wear earplugs, which were provided by the recording team, who also wore them. In such cases the post-processing of the RIRs described in Section 2.4.2 was particularly important to restore a sufficient SNR.

The RIR was measured in each location with the loudspeaker positioned in front of the microphone array at 1.3 m on the horizontal plane. Thereby, the microphone array pointed always to 0° and the loudspeaker to the microphone array. The source-receiver distance of 1.3 m was chosen to minimize any potential near-field effects distorting the perceived spatial image of the reproduced sound source and, at the same time, to realize a short-enough distance that is representative for natural conversations (e.g., see [83]).

2.4. Signal Processing for Playback

2.4.1. HOA Realization

The process of encoding the microphone array signals into the multi-channel HOA format and subsequently decoding it into loudspeaker signals is illustrated in Figure 2, and the corresponding mathematical details are given in Appendix A1. The $Q = 62$ signal channels recorded with the microphone array, $s_{q=1,\dots Q}(t)$, are encoded into $K = 31$ HOA signals, $b_{k=1,\dots K}(t)$, by applying a matrix of $K \times Q$ encoding filters $h_{E,kq}$. The encoding filters were derived in the frequency domain by applying the shape-matching

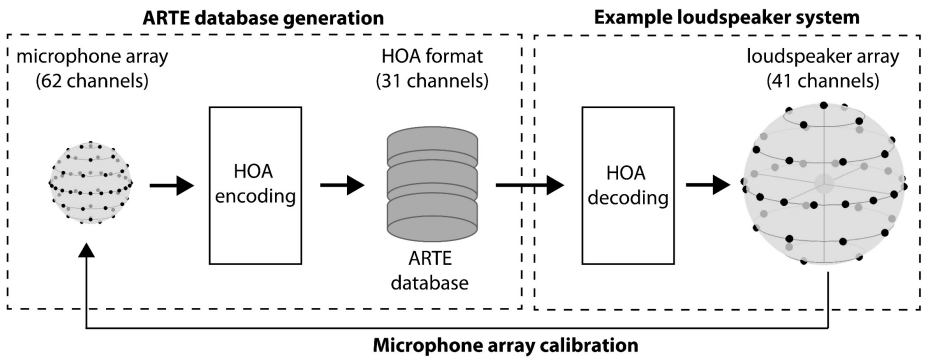

Figure 1. Illustration of the generation and playback of the acoustic scenes contained in the ARTE database.

method [68], which also calibrates the microphone array. In this method, the microphone array is placed in the center of a 3D loudspeaker array, and the impulse responses (IRs) are measured from each loudspeaker to each microphone. The encoding filters are then derived from the measured IRs following the calculations described in the Appendix A1. The finite impulse response (FIR) encoding filters had a length of 2000 samples at a sampling frequency of 44.1 kHz. Limited by the employed loudspeaker array, different HOA orders, $M_{3D} = 4$ and $M_{2D} = 7$, are provided for periphonic (3D) and horizontal-plane only (2D) playback. All spherical harmonic functions up to the degree $m = M_{3D}$ are provided (25 channels in total) as well as all sectorial harmonic functions (i.e., $m = n$, see Equation (A6(in Appendix A1) with degree $M_{3D} < m \leq M_{2D}$ (i.e., 6 additional channels). This results in $K = 31$ HOA channels in total.

Since the shape-matching process inherently involves the frequency response of the loudspeakers of the playback array, they had to be pre-equalized. To equalize the loudspeakers, their individual IRs were measured to an omnidirectional $1/4''$ Type 46BL G.R.A.S. low-noise microphone located in the center of the loudspeaker array, which was powered with a G.R.A.S. CC Supply Type 12 AL and recorded using an RME M-16 analog-to-digital interface. The equalization filters were realized by mixed-phase FIR filters with a length of 2048 samples, which equalized the anechoic loudspeaker responses at 40-10,000 Hz. Applying pre-equalized loudspeaker signals ensured that the final HOA signals provided in the ARTE database are independent of the playback loudspeakers.

As illustrated in Figure 2, the resulting HOA signals are weighted, using a matrix of $K \times G$ decoding gains g_D, and summed up into G loudspeaker signals $l_{g=1,\ldots G}(t)$. The mathematical details of the decoding process are given in Appendix A1. Since this decoding process depends on the specific layout of the playback loudspeaker array, decoding gains cannot be provided here. Instead, a MATLAB (version R2018a) function is provided in the database that takes the loudspeaker locations as input and calculates the gains g_D. With respect to the present study, the example array of $G = 41$ loudspeakers shown in Figure 1 was employed. Given the non-regular layout with an increased number of loudspeakers in the horizontal plane, the mixed-

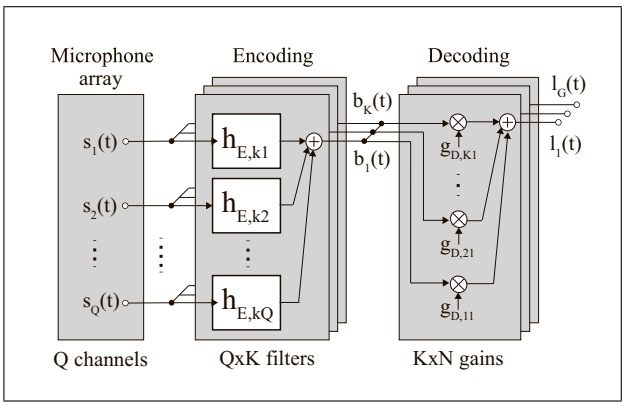

Figure 2. Signal flow diagram of the applied HOA encoding and decoding process. The principal processing is described in the main text and the mathematical details are given in Appendix A1.

order ambisonics method was used (e.g., [30, 53]), also with $M_{3D} = 4$ and $M_{2D} = 7$, utilizing all $K = 31$ HOA channels. In this case, using the same loudspeaker array as in the shape-matching process ensured that the entire sound reproduction system was calibrated such that any recorded scene would be automatically reproduced at its original sound pressure level. For calibrating any arbitrary loudspeaker array for playback, a diffuse speech-shaped noise is provided (see Table I). The sensitivity of the playback system should be adjusted such that the diffuse noise produces a level of 70 dBSPL in the center of the array. The resulting calibration gain can then be applied to any acoustic scene of the ARTE database to reproduce their original sound pressure level.

Finally, it should be noted that depending on the layout of the playback loudspeaker array, not all provided HOA channels may be usable [82, Eqs. 22 and 30]. As a rough guide for regular loudspeaker arrays, at least $N \geq (M_{3D} + 1)^2$ or $N \geq 2 \cdot M_{2D} + 1$ loudspeakers are required for periphonic (3D) or horizontal-plane only (2D) playback. The number of loudspeakers N has to be increased for non-regular loudspeaker arrays. For "massive loudspeaker arrays" the number of applied loudspeakers should be limited to avoid sound coloration [77]. For "reasonable" 2D sound reproduction, only the sectorial HOA channels (i.e., for which $m = n$, in Equation (A6), Appendix A1) may be used in the decoding process. How-

ever, such (or any other) mapping of a 3D scene into 2D will always introduce spatial distortions, amongst other artifacts.

2.4.2. Processing of RIRs

In addition to the recordings of the acoustic scenes, multichannel RIRs are provided in the ARTE database for all the corresponding environments, and are saved as 31-channel HOA signals. All the RIRs were truncated just before the response "disappeared" in the noise floor of the measurement. Since not all recording locations could be accessed during quiet times, some RIRs were contaminated with substantial ambient noise. For some environments, this resulted in rather short RIRs after truncation. Since the RIRs were measured for rather close source-receiver distances, this is not necessarily a problem in applications where substantial background noise (i.e., the acoustic scenes) is presented to the listener and thereby masks the late reverberation of the target speech. However, this may be a problem if the unprocessed RIRs are used to realize a reverberant sound source in quiet.

In some applications it may be useful to enhance the "directionality" of the RIRs (e.g., [70]). For such cases, a second version of all RIRs is provided in the ARTE database, in which the direct sound component was separated from the rest of the RIR and can be expressed as a single-channel RIR. The direct sound component is then given on a separate (non-HOA) channel, which can be presented from a single loudspeaker located (roughly) in the original direction (and distance) of the direct sound while the rest of the RIR can be presented via the loudspeaker array. The direct sound component was separated here by applying a one-sided Hanning time-window to the 31 HOA signals with a frequency-dependent duration of $D = 2/f$, which was limited to the interval $0.002s \leq D \leq 0.01s$. The inverse window (i.e., flipped with 50% overlap) was applied to the reverberant part of the RIR such that an addition of both parts would sum up to the original RIR. The direct sound component was then given by the omnidirectional (the zeroth order) HOA channel. Even though this RIR enhancement can be useful in some applications, it should be noted that it may affect the realism or ecological validity of the reproduced sound field.

2.4.3. Conversion for Binaural Playback

The ARTE database also contains binaural versions of all the recorded acoustic scenes. The binaural signals are mainly provided so that an interested user can get a first impression of the different scenes by listening to them via headphones. These were generated by measuring the head-related transfer functions (HRTFs) for all 41 loudspeakers of the playback array, shown in Figure 1, to the two microphones inside the ears of a Brüel & Kjær type 4128C Head and Torso Simulator (HATS), which describes an array of 41 Tannoy V8 concentric loudspeakers installed within an anechoic chamber. Within this array, the loudspeakers are arranged symmetrically in rings on a sphere with a radius of 1.85 m. Sixteen loudspeakers are mounted on the horizontal plane (0° elevation). Additional sixteen loudspeakers are mounted at ±30° elevation (eight each) and eight at ±60° (four each). One loudspeaker is hung directly above the listener's head. The decoded loudspeaker signals for each acoustic scene were then convolved with the corresponding HRTFs and summed up separately for the left and right ears to form the binaural signals. Additional diffuse field binaural versions of the scenes were obtained by removing the ear canal response from the HATS binaural recordings, through equalization of the ear-drum responses to those of an omnidirectional microphone in diffuse noise field.

2.5. Database Delivery and File Format

The ARTE database requires about 10 GB of storage and is available online at

https://doi.org/10.5281/zenodo.2261632,

where all the necessary information about the implementation, the calibration and playback, and the acoustics of the scenes is available. The documentation about the supplied MATLAB (version R2018a) functions are common to all scenes, and specific examples are given about processes such as applying the HOA decoding to the recordings. Each scene in the ARTE database has its own directory that contains associated data that can be downloaded separately: the 31-channel HOA WAV version of the recording, the binaural WAV version, and a PDF file containing the acoustic parameters of the scene (see Section 3.2 below). Note that the original raw 62-channel recordings are excluded from the database.

3. Acoustic analysis of the database

3.1. Scene Overview

A description of the current set of environments is provided in Table I, along with their mean sound pressure levels in dBA and dB SPL. With the exception of the Train Station scene that opened with a very dominant announcement, and the Street / Balcony scene that had noticeable traffic fluctuations, the excerpts sounded rather consistent over their entire duration, while still revealing the acoustic diversity of the particular scene. The consistent behavior is particularly important for applications in speech-in-noise tests, where major level fluctuations would significantly reduce the test-retest reliability. The scenes were also screened for inappropriate language, as well as for any words that could identify a recorded person or reveal critical (or potentially confidential) information. All excerpts were carefully scrutinized by the authors, to ensure that no recording and HOA processing artifacts were audible. Fade in and fade out (0.5 s long) were applied to the start and end of each recording, respectively, to provide smooth reproduction.

Several environments require special mention. Two church scenes are included, which were both excerpted from the same recording, which was done in identical conditions, but at different times around the service. The two

recordings were made inside a rather small church with uncharacteristically low reverberation, and mainly differed by their level of conversation noise. The Living Room scene contains a sequence of television advertisements that were not recorded in situ. Instead, the IRs of the television set loudspeakers (stereo) were recorded with the HOA microphone array. Random Australian television advertisements were then recorded offline and convolved with the IRs and mixed with the ambient noise of the living room to obtain the final scene. Adding the television sound during post-processing provided some freedom in selecting a generic program and ensured that the presentation level was reasonable, independent of the ambient noise. Finally, a speech-weighted diffuse noise scene (Scene 6 in Table I) is included with an arbitrarily chosen sound pressure level of 70 dB SPL. This scene was artificially generated and recorded inside a 3D loudspeaker array (see section 2.4.2) and is mainly provided for calibrating the applied loudspeaker playback system.

Additional excerpts for the existing as well as for new environments will be appended to the database in the future.

3.2. Derived Scene Acoustic Data

With reference to the third goal of this study (Section 1.4), the results of an acoustic analysis are provided in the ARTE database to allow the potential user to make an informed decision on which scenes to select for their given application, allow comparisons with acoustic scenes provided by other existing databases (see Section 1.2), and help to interpret results when applied in subjective experiments. The acoustic analysis was performed by simulating the reproduction of the recorded scenes through the 41 channel loudspeaker array shown in Figure 1. The simulations were realized by first measuring the IRs between each of the 41 loudspeakers to a G.R.A.S. Type 46BL omnidirectional $1/4''$ microphone located at the center of the loudspeaker array. Then, the decoded HOA loudspeaker signals were convolved with those IRs and summed. The advantage of using this simulation was that it enabled offline estimation of the acoustic parameters and thereby avoided remeasuring each time a new excerpt of an acoustic scene was selected or processed. Details of the acoustic analysis are given below, along with exemplary results for two representative scenes. The ARTE database contains results for all acoustic scenes and will be extended to any new scene that will be added in the future.

3.2.1. Sound pressure level

From the simulated omnidirectional recordings, the unweighted and A-weighted sound pressure levels were calculated for all the different scenes contained in the ARTE database. The results are summarized in Table I. The scenes provide a broad range of levels of about 30 dB in 1-4 dB steps.

3.2.2. Reverberation time

The reverberation times (T_{30}) associated with all the acoustic scenes that are contained in the ARTE database were derived from the simulated omnidirectional RIRs following the process described in ISO 3382-2 [41]. The results are summarized in Table I. In the artificial Diffuse Noise and the Street / Balcony scenes the concept of reverberation time does not apply and no values are provided.

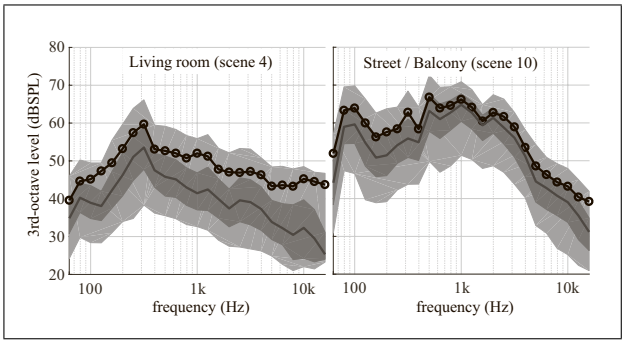

Figure 3. Example power spectra in third-octaves for the Living Room scene (4) and the Street / Balcony scene (10). The long-term power spectra are shown by the solid lines with circles. The median values for a short-term frequency analysis are shown by the solid gray lines, and the corresponding 25th and 75th, as well as the 5th and 95th percentiles are shown by the dark gray and light gray areas, respectively.

3.2.3. Spectrum

To characterize the spectral behavior of the different acoustic scenes, the power spectra in third-octave bands were calculated from the simulated omnidirectional recordings and two examples are shown in Figure 3: the Living Room (scene 4) and the Street / Balcony (scene 10). In addition to a long-term frequency analysis, shown by the solid lines and circles, a short-term frequency analysis was performed using a 20-ms long van Hann windows with 50% overlap. The resulting median values are shown by the solid gray lines, and the 25th and 75th, as well as the 5th and 95th percentiles are shown by the dark gray and light gray areas, respectively. The two example spectra do not only show differences in overall power, but the Street / Balcony scene (right panel) has also relatively more low-frequency power than the Living Room scene (left panel). The percentile plots indicate that both recordings contain substantial level fluctuations of more than 20 dB for the Street / Balcony and more than 30 dB for the Living Room scene.

3.2.4. Temporal envelope and modulation spectrum

To illustrate the temporal behavior of the different acoustic scenes, the temporally smoothed envelope of the simulated omnidirectional recordings was calculated by applying an A-weighting filter to the waveforms, followed by squaring, low-pass filtering at 16 Hz with a 4th-order Butterworth infinite impulse response (IIR) filter, and then taking the square root. The resulting envelope is shown in Figure 4 for the same two example scenes as used above. The modulation spectrum was derived by applying a simplified, frequency-independent version of the processing described by [44]. The waveforms were band-pass filtered

Table I. The ARTE recorded scenes, ordered by unweighted increasing sound pressure level. The scenes are numbered for convenience and their nominal title is provided (see Section 4), along with a general description of what happens in the scenes. Sound pressure level is: unweighted (dB SPL) and A-weighted (dBA). The reverberation time T_{30} was estimated from the RIRs (see Section 3.2.2). *Both in the Café (2) and the Train station scenes, late reflections that arrived from the remote part of the structure were truncated in the T_{30} calculation, which may have resulted in underestimating it, giving drier impression of the sound.

	Scene Name	Description	SPL (dB)	SPL (dBA)	T_{30}(s)
1	Library	University study area in the main library, off-peak hours, quiet	53.0	46.1	0.6
2	Office	Open space office, people typing, chatting and talking on the phone	56.7	51.4	0.2
3	Church (1)	Small church space, people entering and chatting quietly before service	60.5	54.7	1.2
4	Living Room	Living room with access to kitchen in the back, loud television and sounds from the kitchen	63.3	58.7	0.2
5	Church (2)	Same as 3, but busier and louder conversations (1.5 minutes)	65.9	60.9	1.2
6	Diffuse Noise	Speech weighted broadband diffuse sound field at 70 dB SPL	70	65.9	N/A
7	Café (1)	Indoor café at medium occupancy	71.0	67.3	1.1
8	Café (2)	Indoor (company) café at medium occupancy before lunch, next to the wall	71.7	66.2	1.1*
9	Dinner Party	Small room with 8 people chatting over the table with background music	72.8	68.7	0.4
10	Street / Balcony	Apartment balcony over a busy arterial road; Mainly traffic noise with some noise from within the apartment	74.5	71.1	N/A
11	Train Station	Sydney Central, main concourse – large space, open to the platforms with people walking at peak hour; Loud announcement and train sounds	77.1	73.6	1.0*
12	Food Court (1)	Busy university food court	78.2	74.9	0.9
13	Food Court (2)	Very noisy food court in a shopping mall during lunch	79.6	76.7	1.0

by an A-weighting filter, squared, analyzed by a modulation filter bank with one-octave wide band-pass filters, and normalized by the total power of the A-weighted waveform. For plotting purposes, the spacing of the modulation filters was 0.1 Hz and the spectrum was normalized to its maximum value within the modulation frequency range shown in Figure 4. It can be seen that the temporal envelope of the Living Room scene (top-left panel) contains more temporal fluctuations than the Street / Balcony scene (top right panel), which contains more continuous noise. This is further evident when considering the corresponding amplitude modulation spectra shown in the bottom panels of Figure 4. The Living Room scene contains strong temporal modulations with a peak frequency of around 2.8 Hz, which mainly stems from the voice presented from the television. The Street / Balcony scene exhibits very strong low-frequency modulations and shows a minimum at around 5 Hz.

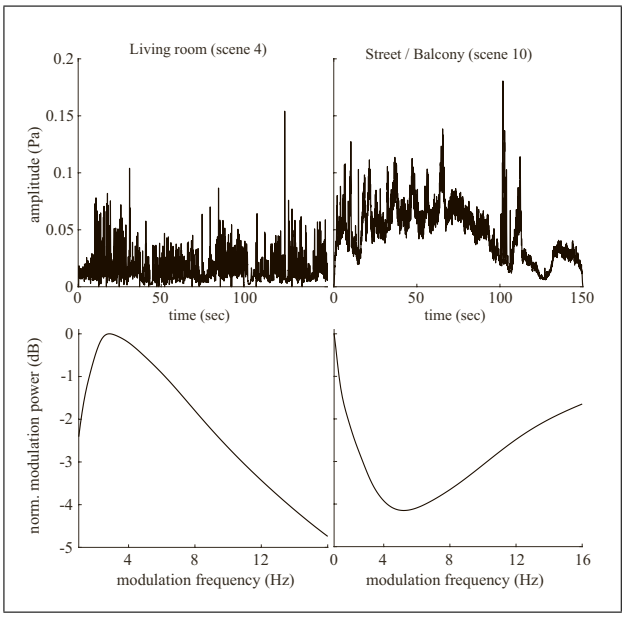

Figure 4. Temporal analysis for the Living Room scene (4) and the Street / Balcony scene (10). The upper panels show the temporally smoothed envelope of the signals recorded by an omnidirectional microphone, and the bottom panels the corresponding amplitude modulation spectra.

3.2.5. Directional characteristics

To illustrate the directional characteristics of the different acoustic scenes, the A-weighted (RMS) sound pressure level of the signals presented by the 16 horizontal loudspeakers of the array shown in Figure 1 were calculated, and are shown in Figure 5 for the two scenes, as a function of loudspeaker direction. These directivity plots

were normalized for each scene separately to the maximum occurring sound pressure level. In the Living Room scene, the main energy arrives from the front left (−45°) and front right (+45°) direction, which corresponds with the location of the loudspeakers of the television. Additionally, there is increased energy coming from the back where the kitchen noise came from. The Street / Balcony scene is less directional, highlights that the microphone array was not directed straight to the road but slightly turned to the left. The energy from behind-left refers to a major reflection from a back wall.

4. Perceptual Evaluation

The overall goal of the the ARTE database is to share stimuli that are significantly more realistic than the stimuli currently used in hearing research (Section 1.4). The accuracy of the applied HOA sound reproduction method has already been discussed by [69] and [71] using different acoustic measures (see also Section 2.4.1), and in [70] the effect on speech intelligibility performance has been evaluated in in hearing impaired listeners with directional hearing aids. Here, a perceptual evaluation of the acoustic scenes described in Table I was performed, with the primary goal of understanding how well the reproduced scenes represent the nominal (original) scenes. For example, does the recorded café scene actually come across as a café? Making this association may not be trivial for the listener without any verbal background, visual context, or prior exposure to the scenes. The questions of this task were part of a larger survey concerning complex acoustic scenes, which will be published elsewhere. However, for many applications the ability to correctly identify the exact scene may not be important, but rather to understand the more general scene category that subjects associate with it. An additional goal of the subjective evaluation was therefore to record the alternative associations that the subjects reported for the different scenes of the ARTE database. This combined information will then help researchers to select environments for their specific application and aid future research on finding more general relevant scene categories as well as auditory scene recognition and analysis.

4.1. Methods

A group of 66 subjects (18 male, 48 female), aged between 19-64 (mean age 29.3 years) participated. Pure tone audiograms were measured for all participants: 50 had normal hearing (≤20 dBHL), twelve had slight hearing losses (20-25 dBHL), and four had mild losses (25-35 dBHL). The subjects received either a small gratuity or course credit for their participation.

All signals were generated on a PC with an RME MADI sound card connected to two RME 32-channel digital-to-analog converters (M-32). These fed 11 Yamaha XM4180 power amplifiers that drove the 41-channel HOA loudspeaker array described in Section 2.4.3 and located in the anechoic chamber of the Australian Hearing Hub, Macquarie University, Australia. The subjects were seated on a height-adjustable chair with their head located in the acoustic center of the HOA loudspeaker array. Fourteen different environments were presented in two parts comprising all the 13 scenes described in Table I, whereby the Diffuse Noise (scene 6) was included as a reference condition and was repeated at two levels: 60 and 70 dB SPL. The first presentation was a training and familiarization round with Café (2) environment (see Table I). Then, a randomized sequence of seven additional environments was presented. Following a mandatory break, the test resumed with one repeated environment out of the seven and a randomized sequence of the remaining six environments. The entire test lasted between 1.5 and 3 hours (including tasks unreported here).

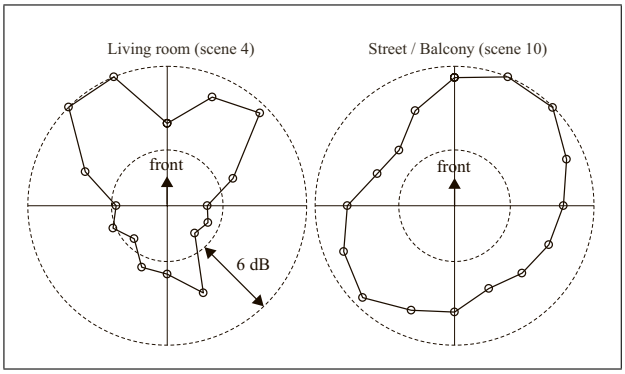

Figure 5. Directivity plots in the horizontal plane for the Living Room scene (4) and the Street / Balcony scene (10).

Test participants were asked to: (i) indicate whether the scene they heard takes place indoors, outdoors, or in a combined space, (ii) try and identify the scene as an open question, (iii) rate how realistic the scenes sounded to them on a scale between 0 to 10 from completely artificial to completely realistic. The questions were answered using paper and pen, and are given in Appendix A2.

All scenes were two minutes long and played in a continuous loop. The subjects were instructed to listen carefully to the acoustic scenes before answering the questions. The experimenter could monitor the participant via a video camera and talk-back microphone system in the control room to provide assistance, if required, during the test and to change the scene when requested.

4.2. Results

The subjective response about the type of scene (indoors, outdoors, combined space) are listed in three columns of Table II, indicating the various confusions made. For the eight quietest scenes, subjects correctly identified the type of environment they listened to, at a rate of 89% or more for the eight quietest scenes. In the Street / Balcony scene, which was classified as combined indoor-outdoor space, most responses (73.8%) indicated that it was completely outdoors, despite some indoor sounds that were part of it. Similarly, the Train Station scene was only correctly identified as a combined indoor-outdoor scene by 52.3% of the subjects, with split responses for indoor or outdoor. Interestingly, indoor identification rates dropped for the loudest scenes (80.7% and 66.7% for Food Courts 1 and 2, re-

Table II. Subjects' place identification through listening to 14 scenes. [a]The number of subjects per scene. Note that the church and diffuse scenes, only the first (unbiased) occurrences were counted. [b]Type of space identification, with correct answers (%) are in boldface: **In**doors, **Out**doors, or **Com**bined space. The question is not strictly valid for the diffuse scenes, but the typical confusions are shown. [c]The strict identification rate (%) estimated by strict verbal equivalence between the written guess by the subject and the known scene location. [d]Lenient identification rate (%) and labels when they were considered close enough to the known scene location. [e]Common alternative labels that may be adapted in future experiments. [f]The mean realism ratings (0-10) with confidence interval of two standard errors. See text for further details.

	Scene Name	N[a]	In[b]	Out[b]	Com[b]	Str ID[c]	Len. ID[d]	Lenient labels[d]	Alternative Labels[e]	Realism[f]
1	Library	65	**93.8**	1.5	0.	60.0	64.6	Study Area (4.6)	Classroom (9.2)0	8.8 ± 0.4
2	Office	65	**95.4**	0.0	0.0	86.2	87.7	"Student Connect" (1.5)	Waiting room (4.6)	8.4 ± 0.4
*	Diffuse Noise (1)	32	40.6	46.9	9.4	0.0	6.3	White noise (6.3)	Heavy rain from inside the house (34.4), Waterfall (15.6), Rain (12.5)	4.6 ± 0.8
3	Church (1)	32	**93.8**	0.0	6.3	9.4	9.4	N/A	Nursing home (15.6), Community hall (9.4)	7 ± 0.6
4	Living Room	66	**89.4**	0.0	6.1	34.8	86.3	Kitchen (39.4), Home (12.1)		8.9 ± 0.4
5	Church (2)	33	**93.9**	3.0	3.0	0.0	0.0	N/A	Social gathering (15.2), Communal Area (12.1)	7.1 ± 0.6
6	Diffuse Noise (2)	33	18.2	66.7	12.1	0.0	9.1	White Noise (9.1)	Rain falling (18.2), Home when raining (15.2), Waterfall (15.2)	4.8 ± 0.8
7	Café (1)	66	**93.9**	0.0	4.5	84.8	95.4	Restaurant (10.6), Cafeteria (3.0)	N/A	8.7 ± 0.3
8	Café (2)	66	**92.4**	0.0	6.1	83.3	95.4	Restaurant (9.1), Cafeteria (3.0)	N/A	8.5 ± 0.4
9	Dinner Party	65	**96.9**	0.0	0.0	26.2	35.4	Living Room (7.7), House (1.5)	Bar (20), Restaurant (18.5), Cafe (13.8), Living room (7.7)	7.9 ± 0.5
10	Street / Balcony	65	0.0	73.8	20.0	6.2	29.3	Roadside cafe (10.8), Street near cafe (4.6), Farm next to street (1.5), Backyard (1.5)	Busy road (32.3), Roadside (18.5)	8.4 ± 0.5
11	Train Station	65	26.2	16.9	**52.3**	100	100	N/A	N/A	8.9 ± 0.3
12	Food Court (1)	66	**80.3**	4.5	13.6	33.3	45.4	School canteen (7.6), Cafeteria (4.5)	Cafe (30.3), Restaurant (18.2)	8.5 ± 0.5
13	Food Court (2)	66	**66.7**	7.6	21.2	33.3	36.3	Mall (1.5), Shopping center (1.5)	Restaurant (16.7), Cafe (10.6), Pub (10.6), Bar (7.6)	8.3 ± 0.4

spectively), possibly because the loud noise masked most of the reverberant cues in these spaces and their contents could be heard also in some outdoor locations. Finally, the diffuse noise scenes, which contained no room cues, were generally associated with outdoor spaces, but at a lower likelihood for the quieter condition (46.9% at 60 dB SPL and 66.7% at 70 dB SPL).

Correct scene identification varied dramatically between scenes (0 to 100%), as listed in the "Strict ID" column of Table II. While the Train Station scene was always successfully identified, the two church scenes were generally confused for something else, or were given a rather vague description (e.g., social gathering). Some of the answers were inaccurate, but were close enough given the

lack of prior knowledge about the scenes, so they were counted as correct. For example, cafeteria and café are very close. This was taken into account in the alternative identification calculation (the "Lenient ID" column of Table II), which accepts more answers as correct. Some identification labels were wrong in the strict sense, but are justifiable and can come across as convincing, especially since they repeated several times. These answers are listed under "Alternative labels" in Table II). Because the scene identity was always revealed after the initial responses (see Appendix A2), many subjects were able to identify correctly the Diffuse Noise as well as the Church scenes when presented a second time (e.g, they identified Church 2 because they had already heard Church 1, which was recorded in the same place, but at another time). To avoid any learning bias for these two scenes, the identification responses were only counted for the first presentation and were omitted when repeated, which is reported in the number N of subjects column of Table II. The most frequent scene identification confusions are also listed in the Table II under the heading of "alternative responses". The subjective response about the type of scene (indoors, outdoors, combined space) are listed in three columns, indicating the various confusions made. Note that the identification questions were asked about the diffuse noise scenes, although they were not strictly answerable, because they were not real places. Three people who associated white noise with these scenes were counted as correct in the lenient identification.

The realism ratings of the scenes exhibit three clusters (Table II, last column): diffuse scenes (mean rating of 4.6-4.8), church scenes (7-7.1), and all the rest (7.9-8.9). The artificial diffuse noise scenes were indeed judged to be more artificial than real, confirming that listeners reacted differentially to these sounds. Additionally, in agreement with the church misidentification, the church scenes were also judged to be less realistic.

4.3. Discussion

Listeners were able to identify correctly a number of scenes, despite the lack of visual cues or additional information. Notably, the train station, the two café and the office scenes were all above 83%, regardless of how the identification was estimated. However, the identification was never perfect and may have depended on the subject's familiarity with the specific environment. Even for a straightforward environment such as a living room, imprecise identifications were common, as the scene is dominated by both loud television and kitchen utensils, so that listeners had to decide where they stand in relation to the different sources. The library had more than 60% identification, which was likely affected by the familiarity of subjects with that particular space, as the recording was done in the university library. The two different food court scenes were misidentified more often than expected. Food Court (1) was recorded in the university itself, which may explain the somewhat higher (lenient) rating of 45.4%, compared to Food Court (2), 36.3%, because it is more familiar at least to the many students that participated in this listening test. However, most confusions of these two scenes indicate that they were correctly associated with noisy food establishments of similar or smaller scale. It is not impossible, for example, that a particularly large pub with a few hundreds of people would sound similar to a crowded food court. Familiarity may have also played a role here, as some people may avoid eating out at food courts altogether. For this reason, some users of the ARTE database may decide that these alternative labels are good enough and can serve as valid scenes, even though they do not adhere to the strict original location. Determining what scene labels should count as 'lenient' or 'alternative' is a somewhat subjective process in itself, but as the complete subjective data is included in Table II, future users of the scenes may reinterpret these labels.

The observed identification rates were only in partial agreement with a previous study [38], in which subjects had to identify stereophonic recordings from 34 locations with an average duration of 10.42 s, and RMS-level matched to 80 dB SPL. For instance, the identification rate was 95% for their train station, similar to the ARTE train station (100%), but only 20% to their office and library recordings, which were 86% and 60%, respectively, in the present study. The differences are difficult to discuss without more details about their recordings, but it is likely that due to their short stimulus presentation, the acoustic scenes did not include enough sound events and acoustic features that are uniquely associated with the particular scenes in order to make a quick judgment.

Identification of the two church scenes was exceptionally low, once the presentation order correction was applied. While almost all listeners could tell that this is some kind of a social setting happening indoors, there were very few cues that disclosed the exact purpose of the social interaction. Some participants commented that, as churchgoers, they found neither the room acoustics (less reverberation than typical for churches; see Section 3.1) nor the conversations sounded like a typical church. It seems warranted that these scenes could be used to represent other social situations but a church. In particular, Church (2) has never been identified as a church, so it may be renamed as "Social Gathering", as it was the most frequent classification given by subjects. Correct identification of the street / balcony scene was also exceptionally difficult for subjects, because of the combined space in which it was recorded. While most of the sound was unmistakable traffic coming from the street, there are some reflections and odd sounds coming from inside the apartment (e.g., dishes and water flowing), which may be difficult to explain without the visual context, especially for lay listeners. This resulted in either incomplete scenes (i.e., "busy road"), or more creative descriptions such as a "road side café". The four respondents whose answers were counted in Table 4.2 identified the scene as being in a house near a busy road. Finally, while the diffuse noise scenes were not meant to sound like any specific place, they were frequently associated with water – either waterfalls or rain.

Identifying the type of space of the recordings (i.e., indoors, outdoors, or combined-space) was generally easier for listeners. The ARTE database currently contains only indoor and two combined space scenes, but unfortunately no outdoor scenes. Most listeners were able to easily identify all indoor scenes, with the exception of the two food courts, where the very loud babble likely masked any room acoustic cues. Also, confusions between combined-space and outdoors were rather common for the two relevant scenes (train station and street/balcony). It is likely that the combined-space option requires a better ability to visualize the scenes and have awareness of room acoustics – something that not all listeners can be expected to have. The unavailability of outdoor scenes did not enable the simulation and subjective evaluation of free-field or natural environments, and thus conclusions from the test cannot be directly extrapolated to such settings based on the available data. Once again, while not strictly correct, it may be justifiable for future users of the ARTE database to treat the Street / Balcony scene as outdoors, given the subjective data of Table II.

The rating of realism turned out to involve a significant degree of ambiguity, which was mainly due to an insufficient explanation of what was meant by "realism". Some subjects referred to the reproduction of the HOA sound system, and other subjects referred to the believability of being in a given scene, once it became known, or to the question if the particular sounds actually represent such a scene. Unfortunately, the final ratings likely refer to a combination of all of these aspects. When asked informally after the test about their overall impression of the sound reproduction, most people found it very real-sounding.

5. General Discussion and Conclusions

While the main aim behind the ARTE database is to provide real-world material in hearing testing, it has been designed in a way that is general enough to be useful in other related fields of research. For example, although the present ARTE database is not intended to serve as a systematic survey of the acoustics in everyday listening environments, it does provide detailed derived acoustic data alongside each recording. Similarly, the level of scene description in ARTE may not match that of soundscape databases, but it may still be suitable for related soundscape as well as ecological acoustics research that requires generating soundscapes under controlled conditions. In room acoustics and digital signal processing, methods based on RIR databases regularly rely on the synthesis of noisy reverberant environments, whereas the ARTE database provides direct recordings of complex every-day environments in addition to RIRs. This results in richer, more complex, and more realistic acoustic scenes, than they are feasible with RIR-based synthesis methods alone (e.g., by being able to capture the acoustic source movement), and may be used for similar applications of improving speech reception through digital signal processing. Finally, in applications such as automatic scene classification, there has been no mention of the absolute levels of the recorded material (except for DEMAND, [3]) – a critical factor in sound perception. The ARTE database is possibly too small to be used in these applications, but it may be used to cross-check the performance of classification algorithms trained on larger, yet uncalibrated, databases.

Although the research questions of the fields of investigation mentioned above – hearing assessment, acoustic communication, soundscapes, room acoustics, scene analysis and automatic scene or source classification – are inevitably different, their requirements for realistic stimuli may be very similar. The release of the ARTE database provides a novel attempt to address several aspects of realistic scenes in a holistic manner. Because of its comprehensive nature, not all features are going to be useful for all researchers. Rather than deterring researchers who may be interested only in a subset of the features, it is important to emphasize that observations, which will be gathered across different studies that investigate various aspects of everyday listening, may become easier to collect and reproduce using the standardized stimuli. Furthermore, it is also hoped that the comprehensive nature of ARTE will encourage its enhancement in the future, with contributions from other researchers.

The HOA sound-field reproduction method that the database is based on is both a disadvantage and an advantage at the current stage of this technology. On the one hand, the strength of using HOA must not be underestimated, as it lends itself to universal deployment using different hardware and software setups. Moreover, deriving binaural, single- or multichannel audio from the HOA recordings is straightforward, even when the complete setup for HOA reproduction is unavailable. It is a proven method to reproduce sound that is subjectively realistic, as was shown in Section 4.2, and is also rich in terms of the instrumental acoustic measures that can be computed from it – spatial, spectral, temporal and dynamic. On the other hand, processing and reproducing the HOA recordings requires technical knowledge that is currently held only by few laboratories around the world. The deployment of the ARTE database will hopefully contribute to making the HOA technology more accessible to the broader research community as a whole, or inspire the development of simpler technologies that could eventually supersede it. Even if this is where the future lies, it would be necessary to investigate to what level of precision realistic virtual scenes have to be reproduced in the laboratory, in order to obtain observations that are relevant to the real world. More research using stimuli such as provided in the ARTE database will be needed to bridge the wide gap between realistic listening environments and the traditional stimuli in psychoacoustics.

The selection of everyday scenes that presently populates the database is not universal for several reasons. The environments were all recorded indoors or in combined indoor-outdoor spaces – all in urban settings of a large, Australian-English speaking metropolis. Even within the limited population tested, large variations were observed

with regards to how well the scenes are recognized. The success rate of correctly recognizing or identifying scenes may degrade if tested on populations of other cities, in other countries, both in the developed and the developing world. For example, the over-represented cases in the database of food courts may be irrelevant to many people in the countryside, as are the church scenes for the many non-Christians worldwide. All possible applications for research that were indicated above – hearing impairment rehabilitation most notably – should not be bounded geographically or culturally to any specific region. Therefore, more universally applicable research may require adding more scenes to the database that pertain to broader populations, possibly through the inclusion of recordings set in non-English speaking cultures. It will be critical also to add outdoor scenes to the database, which at the moment are completely lacking. However, special care will have to be taken with regards to wind noise, which corrupted several of the recordings that were originally intended for ARTE. Even though wind protectors may be used to reduce wind noise, their benefit for the rather high HOA orders that were recorded here is very limited. Instead, it is highly recommended to obtain wind condition forecasts and select the least windy hours of the day. Finally, some locations may be made somewhat redundant through comparative findings from soundscape studies, but this remains to be seen. The ARTE database is purposely designed in a transparent way to encourage other members of the research community to expand it in the future, or to connect it to their own databases.

In some spaces, the number of people may have significantly altered the reverberation that was captured during the recording, compared to that of the RIR measured without them. This discrepancy may be noticeable in circumstances where the recorded scene and a derived stimulus convolved with the respective RIR are mixed, but will have to be examined according to the particular sounds and uses in question.

Appendix

A1. A Higher-order Ambisonics

The applied higher-order ambisonics (HOA) encoding and decoding process followed closely the methods described in [71] and is briefly summarized below. This approach employs a spherical coordinate system, in which the elevation angle φ is measured from the horizontal plane, and the azimuth angle θ increases counterclockwise from the x-axis to the y-axis as seen from positive z-axis.

A1.1. HOA encoding

As illustrated in Figure 2, to encode the Q signal channels recorded with the 3D microphone array described in Section 2.4.1 into the K HOA signals, $b_k(t)$, provided by the ARTE database for each acoustic scene, the microphone signals, $s_q(t)$, are convolved with a matrix of $K \times Q$ encoding filters, $h_{E,kq}(t)$, as given by

$$b_k = \sum_{q=1}^{Q} h_{E,kq}(t) * s_q(t), \tag{A1}$$

with $*$ being the convolution operation. The shape-matching method (e.g., [53]) was applied to derive the encoding filters. In brief, the microphone array was placed in the center of the 3D loudspeaker array shown in Figure 2, and the transfer functions, $H_{P,gq}(\mathrm{j}\omega)$, were measured between all G loudspeakers to all Q microphones. Using these measurements, the transfer functions, $H_{E,kq}(\mathrm{j}\omega)$, of the encoding filters were then derived by

$$\boldsymbol{E} = \boldsymbol{Y}\boldsymbol{P}^H(\boldsymbol{P}\boldsymbol{P}^H + \lambda\boldsymbol{I})^{-1}, \tag{A2}$$

with

$$\boldsymbol{E} = \begin{pmatrix} H_{E,11}(\mathrm{j}\omega) & \cdots & H_{E,1Q}(\mathrm{j}\omega) \\ \vdots & H_{E,kq}(\mathrm{j}\omega) & \vdots \\ H_{E,K1}(\mathrm{j}\omega) & \cdots & H_{E,KQ}(\mathrm{j}\omega). \end{pmatrix} \tag{A3}$$

$\omega = 2\pi f$, f is the frequency in Hz, λ is a regularization parameter (here $\lambda = 0.4$), $\boldsymbol{I}$ is the identity matrix of size Q, $\boldsymbol{P}$ is the matrix of transfer functions measured between the G loudspeakers and the Q microphones, i.e.:

$$\boldsymbol{P} = \begin{pmatrix} H_{P,11}(\mathrm{j}\omega) & \cdots & H_{P,1G}(\mathrm{j}\omega) \\ \vdots & H_{P,qg}(\mathrm{j}\omega) & \vdots \\ H_{P,Q1}(\mathrm{j}\omega) & \cdots & H_{P,QG}(\mathrm{j}\omega) \end{pmatrix}. \tag{A4}$$

$\boldsymbol{P}^H$ is its Hermitian transpose, and $\boldsymbol{Y}$ is the matrix of the K ideal, real-valued, and frequency-independent spherical harmonic functions (SHF) $Y_k(\theta_g, \varphi_g)$ sampled at the direction of the G loudspeakers with azimuth angle θ_g and elevation angle φ_g, i.e.

$$\boldsymbol{Y} = \begin{pmatrix} Y_1(\theta_1, \varphi_1) & \cdots & Y_K(\theta_1, \varphi_1) \\ \vdots & Y_k(\theta_g, \varphi_g) & \vdots \\ Y_1(\theta_G, \varphi_G) & \cdots & Y_K(\theta_G, \varphi_G) \end{pmatrix}, \tag{A5}$$

whereby the SHFs were ordered (from $1 \le k \le K$) according to Table A1 and defined by

$$Y_{mn}^{\sigma}(\theta, \varphi) = \sqrt{(2m+1)(2-\delta_{0,n})\frac{(m-n)!}{(m+n)!}} \cdot P_{mn}(\sin\varphi) \cdot \begin{cases} \cos n\theta & \text{if } \sigma = +1, \\ \sin n\theta & \text{if } \sigma = -1, \\ 1 & \text{if } n = 0, \end{cases} \tag{A6}$$

where n and m are the order and degree of the SHFs, respectively, $\delta_{0,n}$ is the Kronecker delta, equal to 1 for $n = 0$ and 0 otherwise, and P_{mn} are the associated Legendre functions [27, A.2.2]. Note here that the applied real-valued SHFs comes as pairs of functions as expressed by $\sigma = \pm 1$, which transfer into the different horizontal order spatial components (e.g., [73]), as Table A1 indicates. Within the

Table A1. Organization of the $K = 31$ HOA channels.

Channel k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Degree m	0	1	1	1	2	2	2	2	2	3	3	3	3	3	3	3
Order n	0	1	0	1	2	1	0	1	2	3	2	1	0	1	2	3
σ	x	-1	x	1	-1	-1	x	1	1	-1	-1	-1	x	1	1	1

Channel k	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
Degree m	4	4	4	4	4	4	4	4	4	5	5	6	6	7	7
Order n	4	3	2	1	0	1	2	3	4	5	5	6	6	7	7
σ	-1	-1	-1	-1	0	1	1	1	1	-1	1	-1	1	-1	1

ARTE database, all SHFs up to a degree of $m = 4$ were considered as well as all sectorial SHFs (i.e., $m = n$) for $4 < m \leq 7$.

Equation (A2(was applied here separately at $L = 1000$ equidistant frequencies between $0 \leq f \leq f_s/2$, at a sampling frequency of $f_s = 44.1\,\text{kHz}$, which provided the one-sided transfer functions of the encoding filters $h_{E,kq}(t)$. The encoding filters were then derived by extending these one-sided transfer functions by their frequency-reversed complex conjugate (i.e., utilizing the identity $H_{E,kq}(-\text{j}\omega) = H_{E,kq}(\text{j}\omega)^*$) and applying the inverse Fourier transform.

A1.2. HOA decoding

As illustrated in Figure 2, the decoding into G loudspeaker signals, $l_g(t)$, is realized by a weighted sum over all K HOA signals, $b_k(t)$, i.e.

$$l_g = \sum_{k=1}^{K} g_{D,kg}(t) * b_k(t). \tag{A7}$$

With the decoding weights $g_{D,kg}$ given by the decoding matrix

$$\begin{aligned} \boldsymbol{D} &= \begin{pmatrix} g_{D,11} & \cdots & g_{D,K1} \\ \vdots & g_{D,kg} & \vdots \\ g_{D,1G} & \cdots & g_{D,KG} \end{pmatrix} \\ &= \boldsymbol{C}^T(\boldsymbol{C}\boldsymbol{C}^T)^{-1} = pinv(\boldsymbol{C}), \end{aligned} \tag{A8}$$

which is the pseudo inverse of the re-encoding matrix:

$$\boldsymbol{C} = \begin{pmatrix} Y_1(\theta_1, \varphi_1) & \cdots & Y_K(\theta_G, \varphi_G) \\ \vdots & Y_k(\theta_g, \varphi_g) & \vdots \\ Y_1(\theta_1, \varphi_1) & \cdots & , Y_K(\theta_G, \varphi_G) \end{pmatrix}. \tag{A9}$$

With $Y_k(\theta_g, \varphi_g)$ being the K SHFs sampled at the direction (azimuth θ_g and elevation φ_g) of the G playback loudspeakers as described above. The order of the HOA channels $k = 1, 2, \ldots K$ is summarized in Table A1.

A2. Subjective Questionnaire

1. Does the situation take place outdoors, indoors, or in a space that combines elements from both (for example, a roofed market)? *Indoors / Outdoors / Combination indoors and outdoor*
2. Can you identify the type of scenario? If so, what is it?

The following question was given separately on another page, as it revealed the correct scene identity to the listener. Imagine you are situated in [SCENE NAME] whose sound environment is being virtually reproduced now. In the following question you will be asked to subjectively rate various aspects of this sound environment. There are no wrong or right answers.

3. How realistic do you find this audio environment? *Completely Artificial (0) – Completely Realistic (10)*

Acknowledgments

The research related to the ARTE database was financially supported by the Oticon foundation as well as by the HEARing CRC, established and supported under the Cooperative Research Centres Program – an initiative of the Australian Government. The authors are grateful to Greg Stewart and Barry Clinch for their ongoing technical support. Thanks to Savanna Jones for helping with subjective data transfer.

Ethics

The behavioral measurements in this study complied with the ethics guidelines of the Australian Hearing Human Research Ethics Committee, approval number AHH REC2016-13.

References

[1] Aachen impulse response database. http://www.iks.rwth-aachen.de/en/research/tools-downloads/databases/aachen-impulse-response-database/ [Online; accessed 05-Apr-2019].

[2] Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses. http://medi.uni-oldenburg.de/hrir/. [Online; accessed 05-Apr-2019].

[3] DEMAND: The Diverse Environments Multi-channel Acoustic Noise Database. https://zenodo.org/record/1227121 [Online; accessed 05-Apr-2019].

[4] DESRA: Database of Environmental Sounds for Research Activities. https://zenodo.org/record/2622626. [Online; accessed 05-Apr-2019].

[5] Detection and Classification of Acoustic Scenes and Events (DCASE) 2017. http://www.cs.tut.fi/sgn/arg/dcase2017/ [Online; accessed 05-Apr-2019].

[6] EigenScape. https://zenodo.org/record/1012809. [Online; accessed 05-Apr-2019].

[7] Listen to the world. http://www.stuffinablank.com/ soundscapes.html [Online; accessed 05-Apr-2019].

[8] MARDY: Multichannel Acoustic Reverberation Database at York. https://www.commsp.ee..ic.ac.uk/∼sap/resources/mardy-multichannel-acoustic-reverberation-database-at-york-database/ [Online; accessed 05-Apr-2019].

[9] The Multichannel Impulse Response Database (MIRD). http://www.eng.biu.ac.il/∼gannot/RIR_DATABASE/ [Online; accessed 05-Apr-2019].

[10] Multiple sources Starkey database. http://www.eng.biu.ac.il/gannot/speech-enhancement/multiple-sources-starkey-database/ [Online; accessed 05-Apr-2019].

[11] The Open Acoustic Impulse Response (OpenAIR) Library. http://www.openairlib.net/. [Online; accessed 15-Sep-2018].

[12] Room impulse response data set. http://isophonics.net/content/room-impulse-response-data-set [Online; accessed 05-Apr-2019].

[13] Sten Axelsson, Mats E Nilsson, Birgitta Berglund: A principal components model of soundscape perception. Journal of the Acoustical Society of America **128**(5) (2010) 2836–2846.

[14] Elizabeth Francis Beach, Megan Gilliver, Warwick Williams: The NOISE (Non-Occupational Incidents, Situations and Events) database: A new research tool. Annals of Leisure Research 16(2) (2013) 149–159.

[15] Elliott H. Berger, Rick Neitzel, Cynthia A. Kladden: Noise Navigator^{TM} sound level database with over 1700 measurement values, version 1.8. Technical Report E-A-R 88-34/HP, Univ. of Michigan, Dept. of Environmental Health Science, Ann Arbor, MI, June 2015.

[16] Virginia Best, Gitte Keidser, Jorg M Buchholz, Katrina Freeston: An examination of speech reception thresholds measured in a simulated reverberant cafeteria environment. International Journal of Audiology **54**(10) (2015) 682–690.

[17] Robert S. Bolia, W. Todd Nelson, Mark A. Ericson, Brian D. Simpson: A speech corpus for multitalker communications research. Journal of the Acoustical Society of America **107**(2) (2000) 1065–1066.

[18] Giovanni Brambilla, Luigi Maffei, Maria Di Gabriele, Veronica Gallo: Merging physical parameters and laboratory subjective ratings for the soundscape assessment of urban squares. Journal of the Acoustical Society of America **134**(1) (2013) 782–790.

[19] Albert S Bregman: Auditory scene analysis: Hearing in complex environments. In: Thinking in Sound: The Cognitive Psychology of Human Audition. S. McAdams, E. Bigand (eds.) Chapter 2, pages 10–36. Oxford University Press, 1993.

[20] Urie Bronfenbrenner: The Ecology of Human Development: Experiments by Nature and Design. Harvard University Press, Cambridge, MA and London, England, 1979.

[21] Adelbert W. Bronkhorst: The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. Acta Acustica united with Acustica **86**(1) (2000) 117–128.

[22] Douglas S Brungart, Nandini Iyer: Better-ear glimpsing efficiency with symmetrically-placed interfering talkers. Journal of the Acoustical Society of America **132**(4) (2012) 2545–2556.

[23] Cynthia L. Compton-Conley, Arlene C. Neuman, Mead .C Killion, Harry Levitt: Performance of directional microphones for hearing aids: real- world versus simulation. Journal of the American Academy of Audiology **15**(6) (2004) 440–455.

[24] Martin Cooke: A glimpsing model of speech perception in noise. Journal of the Acoustical Society of America **119**(3) (2006) 1562–1573.

[25] Mary Cord, Deniz Baskent, Sridhar Kalluri, B. C. Moore: Disparity between clinical assessment and real-world performance of hearing aids. Hearing Review **14**(6) (2007) 22.

[26] Jens Cubick, Torsten Dau: Validation of a virtual sound environment system for testing hearing aids. Acta Acustica united with Acustica **102**(3) (2016) 547–557.

[27] Jérôme Daniel: Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia. PhD thesis, University of Paris VI, 2000.

[28] Jérôme Daniel, Sebastien Moreau, Rozenn Nicol: Further investigations of high-order ambisonics and wavefield synthesis for holophonic sound imaging. 114th Audio Engineering Society Convention, Mar. 22-25, Amsterdam, The Netherlands. AES, 2003.

[29] Nathaniel I. Durlach: Equalization and cancellation theory of binaural masking-level differences. Journal of the Acoustical Society of America **35**(8) (1963) 1206–1218.

[30] Sylvain Favrot, Marton Marschall, Johannes Käsbach, Jörg Buchholz, Tobias Weller: Mixed-order ambisonics recording and playback for improving horizontal directionality. 131st Audio Engineering Society Convention, Oct. 20–23, NY, USA. AES, 2011.

[31] Stuart Gatehouse, Graham Naylor, Clous Elberling: Benefits from hearing aids in relation to the interaction between the user and the environment. International Journal of Audiology **42**(sup1) (2003) 77–85.

[32] Michael A. Gerzon: Periphony: With-height sound reproduction. Journal of the Audio Engineering Society **21**(1) (1973) 2–10.

[33] Robert M. Ghent: A tutorial on complex sound fields for audiometric testing. Journal of the American Academy of Audiology **16**(1) (2005) 18–26.

[34] Theo S. Goverts, Steven H. Colburn: Complex acoustic environments: Recordings, processing, experiments. 41st Annual MidWinter Meeting of the Association for research in otolaryncholology, San Diego, CA. ARO, February, 10–14th 2018.

[35] Marc Ciufo Green, Damian Murphy: Eigenscape: A database of spatial acoustic scene recordings. Applied Sciences **7**(11) (2017) 1204.

[36] Giso Grimm, Stephan Ewert, Volker Hohmann: Evaluation of spatial audio reproduction schemes for application in hearing aid research. Acta Acustica united with Acustica **101**(4) (2015) 842–854.

[37] Brian Gygi, Valeriy Shafiro: Development of the database for environmental sound research and application (DESRA): Design, functionality, and retrieval considerations. EURASIP Journal on Audio, Speech, and Music Processing **2010**(1) (2010) 654914.

[38] Brian Gygi, Valeriy Shafiro: The incongruency advantage for environmental sounds presented in natural auditory scenes. Journal of Experimental Psychology: Human Perception and Performance **37**(2) (2011) 551.

[39] Elior Hadad, Florian Heese, Peter Vary, Sharon Gannot: Multichannel audio database in various acoustic environments. IEEE 14th International Workshop on Acoustic Signal Enhancement (IWAENC) pages 313–317. IEEE, 2014.

[40] Richard W. Harris, David W. Swenson: Effects of reverberation and noise on speech recognition by adults with various amounts of sensorineural hearing impairment. Audiology **29**(6) (1990) 314–321.

[41] International Organization for Standardization: 3382-2, Acoustics-measurement of room acoustic parameters – Part 2: Reverberation time in ordinary rooms. International Organization for Standardization, Brussels, Belgium, 2008.

[42] Niels Soegaard Jensen, Claus Nielsen: Auditory ecology in a group of experienced hearing-aid users: Can knowledge about hearing-aid users' auditory ecology improve their rehabilitation. Hearing Aid Fitting: Proceedings of the 21st Danavox Symposium, pages 235–258, 2005.

[43] Marco Jeub, Magnus Schafer, Peter Vary: A binaural room impulse response database for the evaluation of dereverberation algorithms. 16th International Conference on Digital Signal Processing, 2009, pages 1–5. IEEE, 2009.

[44] Søren Jørgensen, Torsten Dau: Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. Journal of the Acoustical Society of America **130**(3) (2011) 1475–1487.

[45] James M. Kates: Digital Hearing Aids. Plural Publishing, San Diego, CA, 2008.

[46] Hendrik Kayser, Stephan Dieter Ewert, Jörn Anemüller, Thomas Rohdenburg, Volker Hohmann, Birger Kollmeier: Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses. EURASIP Journal on Advances in Signal Processing (2009) 298605.

[47] Gerald Kidd Jr., H. Steven Colburn: Informational masking in speech recognition. In: The Auditory System at the Cocktail Party. William A. Yost, Arthur N. Popper, Richard R. Fay (eds), Springer, 2017, 75–109.

[48] Gerald Kidd Jr., Christine R. Mason, Virginia M. Richards, Frederick J. Gallun, Nathaniel I. Durlach: Informational masking. In: Auditory Perception of Sound Sources. John C. Middlebrooks, Jonathan Z. Simon, Arthur N. Popper, Richard R. Fay (eds), Springer, 2008, 143–189.

[49] Brent C. Kirkwood: Information from impact sounds: normal and impaired hearing. PhD thesis, Technical University of Denmark, 2007.

[50] Mathieu Lavandier, John F Culling: Speech segregation in rooms: Monaural, binaural, and interacting effects of reverberation on target and interferer. Journal of the Acoustical Society of America **123**(4) (2008) 2237–2248.

[51] PerMagnus Lindborg: Psychoacoustic, physical, and perceptual features of restaurants: A field survey in singapore. Applied Acoustics **92** (2015) 47–60.

[52] Nicole Marrone, Christine R Mason, Gerald Kidd Jr.: Evaluating the benefit of hearing aids in solving the cocktail party problem. Trends in Amplification **12**(4) (2008) 300–315.

[53] Márton Marschall, Sylvain Favrot, Jörg Buchholz: Robustness of a mixed-order ambisonics microphone array for sound field reproduction. 132nd Audio Engineering Society Convention, 26–29 April, Budapest, Hungary. AES, 2012.

[54] Sven L. Mattys, Matthew H. Davis, Ann R. Bradlow, Sophie K. Scott: Speech recognition in adverse conditions: A review. Language and Cognitive Processes **27**(7-8) (2012) 953–978.

[55] Josh H. McDermott: The cocktail party problem. Current Biology **19**(22) (2009) R1024–R1027.

[56] Annamaria Mesaros, Toni Heittola, Aleksandr Diment, Benjamin Elizalde, Ankit Shah, Emmanuel Vincent, Bhiksha Raj, Tuomas Virtanen: DCASE 2017 challenge setup: Tasks, datasets and baseline system. DCASE 2017 – Workshop on Detection and Classification of Acoustic Scenes and Events, 2017.

[57] Annamaria Mesaros, Toni Heittola, Dan Ellis: Datasets and evaluation. In: Computational Analysis of Sound Scenes and Events. Tuomas Virtanen, Mark D Plumbley, Dan Ellis (eds) Springer, 2018, 147–179.

[58] Annamaria Mesaros, Toni Heittola, Tuomas Virtanen: TUT database for acoustic scene classification and sound event detection. European Signal Processing Conference (EUSIPCO), 1128–1132. IEEE, 2016.

[59] Pauli Minnaar, Sylvain Favrot, Jorg M. Buchholz: Improving hearing aids through listening tests in a virtual sound environment. The Hearing Journal **63**(10) (2010) 40–42.

[60] Brian C. J. Moore: Cochlear Hearing Loss: Physiological, Psychological and Technical Issues. John Wiley & Sons, 2007.

[61] Brian C. J. Moore: An Introduction to the Psychology of Hearing. Brill, Leiden, Boston, 6th edition, 2013.

[62] Brian C. J. Moore, Roger F. Laurence, David Wright: Improvements in speech intelligibility in quiet and in noise produced by two-channel compression hearing aids. British Journal of Audiology **19**(3) (1985) 175–187.

[63] Swen Müller, Paulo Massarani: Transfer-function measurement with sweeps. Journal of the Audio Engineering Society **49**(6) (2001) 443–471.

[64] Damian T. Murphy, Simon Shelley: OpenAIR: An interactive auralization web resource and database. 129th Audio Engineering Society Convention, 4–7 Nov., San Francisco, CA, USA. AES, 2010.

[65] Graham Naylor: Theoretical issues of validity in the measurement of aided speech reception threshold in noise for comparing nonlinear hearing aid systems. Journal of the American Academy of Audiology **27**(7) (2016) 504–514.

[66] Tobias Neher, Thomas Behrens, Simon Carlile, Craig Jin, Louise Kragelund, Anne Specht Petersen, André van Schaik: Benefit from spatial separation of multiple talkers in bilateral hearing-aid users: Effects of hearing loss, age, and cognition. International Journal of Audiology **48**(11) (2009) 758–774.

[67] Michael Nilsson, Sigfrid D Soli, Jean A Sullivan: Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. Journal of the Acoustical Society of America **95**(2) (1994) 1085–1099.

[68] Chris Oreinos, Jörg Buchholz: Validation of realistic acoustic environments for listening tests using directional hearing aids. 14th International Workshop on Acoustic Signal Enhancement (IWAENC), pages 188–192. IEEE, 2014.

[69] Chris Oreinos, Jörg M. Buchholz: Objective analysis of ambisonics for hearing aid applications: Effect of listener's head, room reverberation, and directional microphones. Journal of the Acoustical Society of America **137**(6) (2015) 3447–3465.

[70] Chris Oreinos, Jörg M. Buchholz: Evaluation of loudspeaker-based virtual sound environments for testing directional hearing aids. Journal of the American Academy of Audiology **27**(7) (2016) 541–556.

[71] Chris Oreinos: Virtual Acoustic Environments for the Evaluation of Hearing Devices. PhD thesis, Macquarie University, 2015. http://hdl.handle.net/1959.14/1269481 [Online; accessed 22-May-2019].

[72] Karl S. Pearsons, Ricarda L. Bennett, Sanford Fidell: Speech levels in various noise environments. Technical Report EPA600/1-77-025, Office of Health and Ecological Effects, Office of Research and Development, US EPA, May 1977.

[73] Mark A. Poletti: Three-dimensional surround sound systems based on spherical harmonics. Journal of the Audio Engineering Society **53**(11) (2005) 1004–1025.

[74] Jerry L. Punch, Randy Robb, Amy H. Shovels: Aided listener preferences in laboratory versus real-world environments. Ear and Hearing **15**(1) (1994) 50–61.

[75] Aasa Skagerstrand, Stefan Stenfelt, Stig Arlinger, Joel Wikstrom: Sounds perceived as annoying by hearing-aid users in their daily soundscape. International Journal of Audiology **53**(4) (2014) 259–269.

[76] Karolina Smeds, Florian Wolters, Martin Rung: Estimation of signal-to-noise ratios in realistic sound scenarios. Journal of the American Academy of Audiology **26**(2) (2015) 183–196.

[77] Audun Solvang: Spectral impairment of two-dimensional higher order ambisonics. Journal of the Audio Engineering Society **56**(4) (2008) 267–279.

[78] Rebecca Stewart, Mark Sandler: Database of omnidirectional and B-format room impulse responses. IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), 2010, pages 165–168. IEEE, 2010.

[79] Joachim Thiemann, Nobutaka Ito, Emmanuel Vincent: The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings. 21st International Congress on Acoustics, 2013.

[80] Barry Truax: World soundscape project database. https://www.sfu.ca/∼truax/wsp.html. [Online; accessed 05-Apr-2019].

[81] Kirsten Carola Wagener, Martin Hansen, Carl Ludvigsen: Recording and classification of the acoustic environment of hearing aid users. Journal of the American Academy of Audiology **19**(4) (2008) 348–370.

[82] Darren B. Ward, Thushara D. Abhayapala: Reproduction of a plane-wave sound field using an array of loudspeakers. IEEE Transactions on Speech and Audio Processing **9**(6) (2001) 697–707.

[83] Adam Weisser, Jörg M Buchholz: Conversational speech levels and signal-to-noise ratios in realistic acoustic conditions. Journal of Acoustical Society of America **145**(1) (2019) 349–360.

[84] Jimi Y. C. Wen, Nikolay D. Gaubitch, Emanuel A. P. Habets, Tony Myatt, Patrick A. Naylor: Evaluation of speech dereverberation algorithms using the MARDY database. Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC), 2006.

[85] Adam Westermann, Jörg M. Buchholz: The effect of nearby maskers in reverberant multi-talker environments. Journal of the Acoustical Society of America **141**(3) (2017) 2214–2223.

[86] William S. Woods, Elior Hadad, Ivo Merks, Buye Xu, Sharon Gannot, Tao Zhang: A real-world recording database for ad hoc microphone arrays. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2015. IEEE, 2015.

[87] Yu-Hsiang Wu, Elizabeth Stangl, Octav Chipara, Syed Shabih Hasan, Anne Welhaven, Jacob Oleson: Characteristics of real-world signal to noise ratios and speech listening situations of older adults with mild to moderate hearing loss. Ear and Hearing **39**(2) (2018) 293–304.

[88] Wei Yang, Jian Kang: Acoustic comfort evaluation in urban open public spaces. Applied Acoustics **66**(2) (2005) 211–229.

[89] William A. Yost, Raymond H. Dye, Stanley Sheft: A simulated "cocktail party" with up to three sound sources. Perception & Psychophysics **58**(7) (1996) 1026–1036.