THE ACE CHALLENGE - CORPUS DESCRIPTION AND PERFORMANCE EVALUATION

J. Eaton, N. D. Gaubitch†, A. H. Moore*, P. A. Naylor**

* Department of Electrical and Electronic Engineering, Imperial College London, UK
† SIP Laboratory, Delft University of Technology, Netherlands

ABSTRACT

Knowledge of the Direct-to-Reverberant Ratio (DRR) and Reverberation Time (T_{60}) can be used to better perform speech and audio processing such as dereverberation. Established methods compute these parameters from measured Acoustic Impulse Responses (AIRs). However, in many practical situations the AIR is not available and the parameters must be estimated non-intrusively directly from noisy speech or audio signals. The Acoustic Characterization of Environments (ACE) Challenge is a competition to identify the most promising non-intrusive DRR and T_{60} estimation methods using real noisy reverberant speech. We describe the ACE corpus comprising multi-channel AIRs, and multi-channel noise including ambient, fan and babble noise recorded in the same environment as the measured AIRs, along with the corresponding DRR and T_{60} measurements. The evaluation methodology is discussed and comparative results are shown.

Index Terms— speech enhancement, speech dereverberation, acoustic impulse response

1. INTRODUCTION

The Acoustic Characterization of Environments (ACE) Challenge is a competition devised to stimulate research in the area of blind acoustic parameter estimation from noisy reverberant speech based on a new noisy reverberant speech corpus. Inspired by Gaubitch *et al.* [1], the aim of the challenge is to determine the state-of-the-art in blind acoustic parameter estimation, and to enable and promote research in this field.

To develop acoustic parameter estimation algorithms robust to noise such as [2], it is necessary to simulate a noisy reverberant environment. Whilst established methods exist for simulating Acoustic Impulse Responses (AIRs) [3], and different types of monaural sensor noise can be added to the signal such as [4], simulating diffuse noise emanating from all parts of a room is problematic. How noise arrives at the microphone is dependent on the characteristics of the room, and there will be a different AIR for every point from which noise emanates.

To address the problem of generating test data to represent realistic noisy environments, a unique multi-channel noisy reverberant speech corpus was created. It comprises AIRs, three types of noise recorded in the same room as the measured AIRs, including babble noise for a set of different rooms, two microphone positions within each room, and five different microphone array configurations. The Reverberation Time (T_{60}) and Direct-to-Reverberant Ratio (DRR) were measured for two sets of microphone positions per room, and these measurements are provided with the corpus. In addition, a set of anechoic recordings of free-running speech are provided. By combining the various elements of anechoic speech, AIRs, and associated noise, realistic rooms can be constructed with a wide range of T_{60}, DRRs, noise types, and Signal-to-Noise Ratios (SNRs). The following signal model was used for each channel of audio:

$$y(n) = x(n) * h(n) + \nu(n), \tag{1}$$

where $y(n)$ is the noisy reverberant speech, $x(n)$ is anechoic speech, $h(n)$ is the AIR, and $\nu(n)$ is additive noise.

The contribution of this paper is to describe the ACE corpus including the novel multi-channel babble noise recorded in the same room as the AIR, and the use of the corpus for the ACE Challenge.

The remainder of the paper is organized as follows: In Section 2, the equipment and configurations used for recording the components of the corpus, along with the methods for determining the T_{60} and DRR are described. In Section 3, the ACE Challenge is described. In Section 4, a summary of the results of the ACE challenge are provided, and in Section 5, conclusions are drawn.

2. ACE CORPUS

2.1. Rooms

Seven different rooms within the Department of Electrical and Engineering at Imperial College London were used to produce the corpus. Table 1 lists the approximate dimensions, mean T_{60}, and mean DRR for each room. Their characteristics are as follows

- Office 1: A small lightly furnished carpeted office containing a table, desk and four chairs
- Office 2: A small furnished carpeted office containing a table, desk, 6 chairs and a bookcase
- Meeting Room 1: A medium sized carpeted meeting room containing a meeting table and 14 chairs
- Meeting Room 2: A large carpeted furnished meeting room containing approximately 30 chairs and 6 tables
- Lecture Room 1: A medium-sized hard-floored furnished lecture room containing approximately 20 tables and 60 chairs
- Lecture Room 2: A large hard floored furnished lecture room containing approximately 35 tables and 1-5 chairs
- Building Lobby: A large irregular open room with coupled spaces including a café, stairwell and staircase. The measurements in Table 1 correspond to the corner area where the recordings were made. The total volume of the space is many times larger. The level of ambient noise was high and included non-stationary sources including the main automatic doors and associated card reader, the lifts and the lift announcements, and users of the building passing the recording environment.

Name	L (m)	W (m)	H (m)	Vol. (m^3)	T_{60} (s)	Mic. pos. 1 DRR min. (dB)	Mic. pos. 1 DRR max. (dB)	Mic. pos. 2 DRR min. (dB)	Mic. pos. 2 DRR max. (dB)
Office 1	4.8	3.3	3.0	47	0.34	−2.7	13	−0.55	6.6
Office 2	5.1	3.2	2.9	48	0.39	−0.44	13	−2.3	9.5
Meeting Room 1	6.6	4.7	3.0	92	0.44	−2.0	11	−3.1	7.6
Meeting Room 2	10.3	9.2	2.6	250	0.37	−2.6	11	1.1	12
Lecture Room 1	6.9	9.7	3.0	200	0.64	−0.82	15	0.87	7.9
Lecture Room 2	13.4	9.2	2.9	360	1.25	−0.37	13	−3.7	6.4
Building Lobby	5.1	4.5	3.2	72	0.65	−0.94	13	−2.5	8.1

Table 1: Room dimensions (approx.), mean T_{60}, and mean DRR across all microphone positions, configurations, and channels.

2.2. Microphone configurations, source and seating positions

For each room two separate microphone positions were used. The recording procedure involved first making an empty room AIR recording and noise recordings. The participants would then arrive. Occupied AIR and noise recordings would be made. The microphones would then be moved to the second position. Further occupied AIR and noise recordings would be made. The participants would then leave the room and further unoccupied AIR and noise recordings would be made. The source position and seating position of the occupants remained the same for all recordings.

Five microphone configurations were used in recordings:

- 2-channel laptop with a microphone spacing of 62 mm (Chromebook Pixel). The microphones are situated in a rubberised slot between the screen glass and the outer casing. The keyboard microphone was not recorded
- 3-channel mobile phone array with the microphones arranged in a right-angled triangle with a base of 45 mm and a side of 100 mm. Channel 1 is at intersection of the base and the hypotenuse, channel 2 is at the other end of the base, and channel 3 is at the other end of the hypotenuse
- 5-channel cruciform with a centre-to-arm distance of 250 mm. The centre microphone is channel 1, whilst the remaining channels 2-5 are arranged clockwise viewed from above
- 8-channel linear array with a spacing of 60 mm. Channel 1 is the leftmost microphone facing the source viewed from above
- 32-channel spherical microphone with a diameter of 84 mm (MH Acoustics Eigenmike). The precise orientation of the microphones is described in [5]

High quality microphones were used. For the 3-, 5- and 8- channel arrays, DPA 4060 miniature omni-directional condenser microphones [6] were used, whilst the Eigenmike comprises 32 individually calibrated professional-grade 14 mm electret pressure microphones embedded in an 84 mm rigid sphere baffle.

The 3-, 5- and 8-channel arrays were recorded using two RME OctaMic preamps with their balanced outputs connected to the balanced-line inputs of two RME FireFace 800 Firewire Audio Interfaces. FireFace Recordings were made using Audacity on a MacBook Pro. Eigenmike recordings were made using Eigenstudio on a second MacBook Pro running Windows XP. The Eigenmike interface and the second FireFace 800 were clock-synchronized to the first FireFace 800 interface. The laptop recordings were not clock-synchronised to the other audio devices.

All recordings were made using a sample rate of 48 kHz and 32-bit depth. Laptop recordings were made using *arecord* in little endian format. This technique therefore preceded any equalization which might be performed in the laptop to compensate for the response of the microphone enclosure.

Figure 1 shows the recording equipment in place in the Building Lobby ready to commence recording.

Figure 1: Equipment ready for the recording session in the Building Lobby before the occupants arrive.

2.3. Noise

Three different noise types (ambient, fan and babble) were recorded in each room for each microphone position. To generate the babble noise, between four to seven people were required to sit in the vicinity of the source and speak continuously for the duration of the noise recording. Talkers were provided with a list of phrases from TIMIT [7], or could bring their own material. In a few cases talkers read from scientific papers. Since changes to the acoustics of the room needed to be kept to a minimum during the recordings to ensure that the noises matched the AIRs, talkers remained in their seated positions for the duration of the recording session in each room. The ambient noise was a recording of the room whilst the occupants remained silent. To create the fan noise, one or two fans were placed in the room. Care was taken to avoid wind noise from the fans reaching the microphones.

For all of the rooms and microphone positions except Office 2, AIRs for the rooms without participants were captured. The effect of removing the talkers is discussed in Section 3. For both the fan and babble noises, there is also ambient noise in the back-

ground since this could not be avoided. Completed noise recordings were time-aligned across all microphone configurations using *sigalign.m* [8].

2.4. AIR measurement

Room impulse responses were captured using a frequency sweep [9] played through a calibrated source and recorded using all microphone configurations simultaneously. The source was a Fostex 6301B Personal Monitor which has a 100 mm driver unit. Recordings of the frequency sweep were convolved with the inverse sweep to obtain the AIRs. Silence at the beginning and end of the raw AIRs was removed. The tail of each AIR was faded down to zero over 10,000 samples once the level fell below -70 dB. The AIRs were analysed for fullband and subband T_{60} and DRR as described below. Overall mean values for T_{60} and DRR for each room are shown in Table 1.

2.5. T_{60} measurement

The Energy Decay Curve (EDC) was computed from the AIR, $h(t)$, from the Schroeder integral [10]

$$\mathrm{EDC}(t) = \int_t^\infty h^2(\tau)\mathrm{d}\tau. \quad (2)$$

The method of [11] was then used to estimate the T_{60}. This was found to be more reliable under all conditions than either the ISO-3382 Reverberation Time (T_{30}) or Reverberation Time (T_{20}) derived T_{60} estimates [12] which tended to over-estimate.

Figure 2 shows T_{60}s calculated from different AIR measurements made within a single room. It can be seen that in the range 1 kHz to 5 kHz, the T_{60} is approximately 100 ms larger for the unoccupied room compared to the occupied room, showing that the participants in the experiment absorbed sound energy and reduced the reverberation in the room as expected and confirms the importance of participants remaining in the room during all recordings.

2.6. DRR measurement

The DRR was estimated using the method of [13] as

$$\mathrm{DRR} = 10\log_{10}\left(\frac{\sum_{n=n_d-n_0}^{n_d+n_0} h^2(n)}{\sum_{n=0}^{n=n_d-n_0} h^2(n) + \sum_{n=n_d+n_0}^{n=\infty} h^2(n)}\right), \quad (3)$$

where the direct path signal arrives at sample time n_d, and $n_0 = 2.5$ ms. A delay of 8 ms represents an additional path difference of approximately 23 mm at 340 m s^{-1}. The location of the direct path was found by convolving the AIR with the equalisation filter for the source and finding the maximum, n_d. Equation (3) was then applied to the unequalized AIR.

2.7. Measurement frequency bands

Measurements were made in fullband and in frequency bands using the ISO standard for preferred frequencies [14], thus band 1 is at 25.1189 Hz, band 7 is at 100 Hz, band 17 is at 1000 Hz, band 27 is at 10 kHz, and band 30 is at 19.953 kHz. The filter bank used was a third-octave 8th order Butterworth design. The filter was designed using the Matlab *fdesign.octave* function with bands per octave set to 3 and filter order set to 8. Centre frequencies were generated using the Matlab *validfrequencies* function.

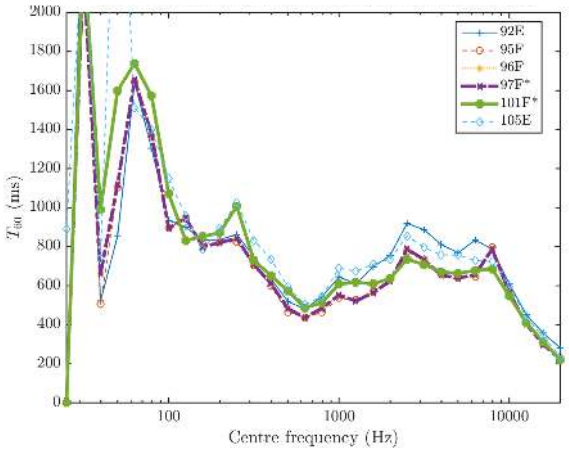

Figure 2: Subband T_{60} measurements for Lecture Room 1. Session codes with E-suffixes indicates that the room was unoccupied.

2.8. Anechoic speech

Two sets of speech were recorded in a single sitting per talker. In set 1, four male talkers were used. The utterances comprise a brief description of where the talker lives, and then a longer description of how the talker gets to work. In set 2, five female and five male talkers were used. The utterances comprise the talker's favourite colour, the town where they live, a description of where they live, a description of how they get to work, and a count from zero to nine. The utterances are in different dialects of international English with a mix of native and non-native English speakers.

Recordings were performed in the anechoic chamber at TU Delft using a B&K 4190-L-001 high quality measurement microphone [15] connected to a B&K NEXUS 2690 conditioning amplifier connected to a RME FireFace 800 audio interface. The speech files were then normalised to give approximately equal loudness measured in Loudness Units Full-Scale (LUFS), and then manually divided into utterances which vary in length by utterance type and speaker. Given a target loudness of -23 LUFS (as defined in European Broadcasting Union (EBU) Recommendation R 128), the aim of the normalization was to ensure all files were within ± 1 LUFS of the target.

3. ACE CHALLENGE COMPARATIVE EVALUATION

The ACE Challenge comprised two phases, the first timed to enable a WASPAA 2015 submission, and the second timed for the ACE proceedings. The ACE Challenge corpus comprised Development (Dev) and Evaluation (Eval) datasets. The purpose of the Dev dataset was to allow participants to review the performance of their algorithms on typical ACE data, supplement their training data, and perform any final training before commencing the challenge. The purpose of the Eval dataset was to provide blind noisy reverberant speech upon which to base the competition. Participants were expected to train their algorithms for T_{60} and DRR estimation against the ground truth values in the Dev data-set using the provided software tools, and then submit blind results to be decoded by

the organizers, for subsequent return of statistics. Participants were expected to present the results in a paper describing their methods.

The Dev dataset comprised noisy reverberant speech recordings from 2 rooms with 2 microphone positions, 4 male talkers from set 1, 2 utterances each, babble, fan, and ambient noise at 0 dB, 10 dB, and 20 dB SNR for all microphone configurations. The noisy reverberant speech files were constructed from anechoic speech from set 1 convolved with the measured AIRs obtained from a given room, with additive noise recorded in the same session, with the same occupants, and with the same source and microphone configuration. A random selection of the noise, either ambient, fan or babble from the same room and microphone configuration and position was then mixed at a predetermined SNRs by equating the active speech power based on ITU-T P.56 [16] with the noise power using the *v_addnoise.m* Voicebox [8] function. Ground truth T_{60} and DRR information was provided to participants for every channel of every microphone position in both fullband and in ISO subbands.

The Eval dataset comprised 5 rooms with 2 microphone positions, 5 male and 5 female talkers from set 2, 5 utterances each, babble, fan and ambient noise at low (-1 dB), medium (12 dB), and high (18 dB) SNRs generated as for the Dev dataset except using anechoic speech from set 2. The audio files for each microphone configuration were numbered in a different pseudorandom permutation to prevent training on the Eval dataset. Both Dev and Eval datasets were resampled to a sample rate of 16 kHz and converted to 16-bit depth. In both Dev and Eval datasets, a further single-channel microphone configuration was included. In the Dev dataset, this used channel 1 of the 8-channel linear array. For the Eval dataset the microphone for the single-channel dataset was selected from channel 1 of the 5-channel cruciform.

4. RESULTS

A selection of the results of the ACE Challenge Phase 1 and 2 single- and two-channel submissions for fullband T_{60} percentage estimation error and DRR estimation error in babble noise at 12 dB SNR are shown in Figs. 3 and 4 respectively. Most submissions in Phase 1 were single-channel fullband estimators with several submissions performing joint T_{60} and DRR estimation, whilst Phase 2 included several multi-channel submissions exploiting the entire corpus. The results show that state-of-the-art single- and two-channel and estimators can estimate T_{60} to within $\approx \pm 25\%$ of the ground truth value, and DRR to within $\approx \pm 2.5$ dB of the ground truth. The results show differing biases and responses to noise resulting from a wide range of approaches, but also that some approaches first published within the scope of the ACE Challenge do not outperform existing methods, and that non-intrusive estimation of T_{60} and DRR is not a solved problem. The full results for the ACE challenge will be provided at a satellite workshop to be held during WASPAA 2015.

5. CONCLUSION

A corpus of multi-channel realistic noisy reverberant speech has been created, initially to support the ACE Challenge, but more widely to support any related research. Seven rooms are provided, each with two source-microphone configurations. For each of the 14 room-source-microphone configurations, 5 different microphone arrays are provided giving up to 50 channels of three different types of noise, along with their AIRs, and T_{60} and DRR measurements. The corpus allows researchers to construct noisy reverberant speech utterances with a range of T_{60}, DRRs, and noises at different SNRs using different microphone arrays in order to evaluate the performance of speech enhancement and speech recognition applications. The effect of occupancy on the T_{60} has also been demonstrated. The database will be made freely available after WASPAA 2015.

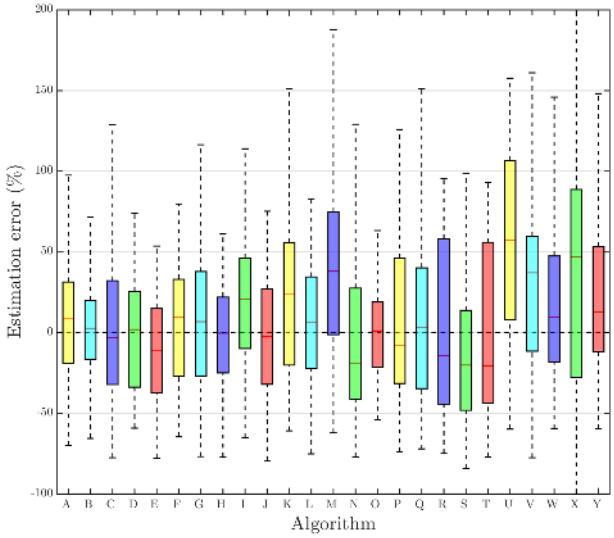

Figure 3: Fullband T_{60} single- and two-channel percentage estimation error in Babble noise by algorithm at 12 dB SNR.

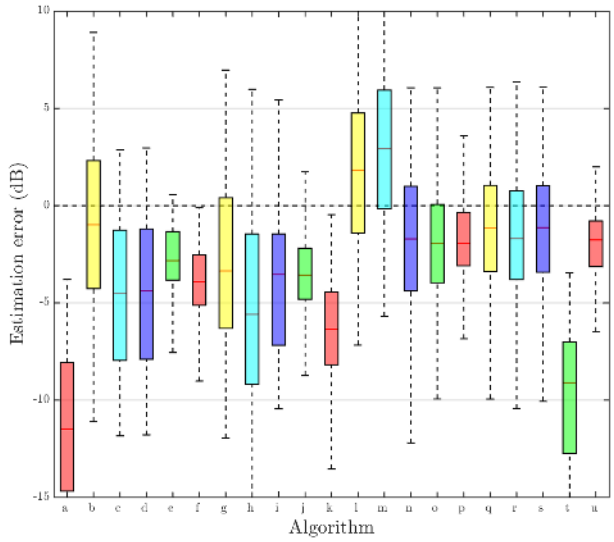

Figure 4: Fullband DRR single- and two-channel estimation error in Babble noise by algorithm at 12 dB SNR.

6. ACKNOWLEDGEMENTS

We wish to thank the following people for participating in the recording sessions: R. Stanton; C. Nelke; L. Lightburn; S. Hafezi; H. Javed; Y. Wang; S. Reynolds; C. Evers; F. Lim; B. Lowe; C. Yiallourides; Z. Jones; J. Murray-Bruce; C. Doire; and A. Stott.

7. REFERENCES

[1] N. D. Gaubitch, H. W. Löllmann, M. Jeub, T. H. Falk, P. A. Naylor, P. Vary, and M. Brookes, "Performance comparison of algorithms for blind reverberation time estimation from speech," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Aachen, Germany, Sept. 2012.

[2] J. Eaton, N. D. Gaubitch, and P. A. Naylor, "Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 161–165.

[3] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[4] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 3, no. 3, pp. 247–251, July 1993.

[5] M. H. Acoustics, "EM32 Eigenmike microphone array release notes (v17.0)," 25 Summit Ave, Summit, NJ 07901, USA, Oct. 2013. [Online]. Available: http://www.mhacoustics.com/sites/default/files/ReleaseNotes.pdf

[6] DPA Microphones, "d:screet 4060 Omnidirectional Microphone, high-sensitivity," Gydevang 42-44 DK-3450 Allerød, Denmark, 2008. [Online]. Available: http://www.dpamicrophones.com/en/products.aspx?c=Item&category=128&item=24035

[7] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, Technical Report, Dec. 1988.

[8] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, 1997–2015.

[9] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Proc. Audio Eng. Soc. (AES) Convention*, no. 108, Feb 2000. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=10211

[10] H. Kuttruff, *Room Acoustics*, 4th ed. London: Taylor & Francis, 2000.

[11] M. Karjalainen, P. Antsalo, A. Mäkivirta, T. Peltonen, and V. Välimäki, "Estimation of modal decay parameters from noisy response measurements," *J. Audio Eng. Soc. (AES)*, vol. 11, pp. 867–878, 2002.

[12] ISO, *ISO-3382 Acoustics - Measurement of the Reverberation Time of Rooms with Reference to Other Acoustical Parameters*, Intl. Org. for Standardization (ISO) Recommendation ISO-3382, May 2009.

[13] S. Mosayyebpour, H. Sheikhzadeh, T. Gulliver, and M. Esmaeili, "Single-microphone LP residual skewness-based inverse filtering of the room impulse response," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1617–1632, July 2012.

[14] ISO, *ISO-266 Acoustics - Preferred frequencies*, Intl. Org. for Standardization (ISO) Recommendation ISO-266, Mar. 1997.

[15] Brüel and Kjær, "4190-L-001 0.5 inch free-field microphone with type 2669-L preamplifier, 3 Hz to 20 kHz, 200 V polarization," Skodsborgvej 307, DK-2850 Nærum, Denmark, 2012. [Online]. Available: http://www.bksv.com/Products/transducers/acoustic/microphones/microphone-preamplifier-combinations/4190-L-1

[16] ITU-T, *Objective Measurement of Active Speech Level*, International Telecommunications Union (ITU-T) Recommendation P.56, Mar. 1993.