Estimation of room acoustic parameters: The ACE Challenge

James Eaton, Student Member, IEEE, Nikolay D. Gaubitch, Alastair H. Moore, Member, IEEE, and Patrick A. Naylor, Senior Member, IEEE.

Abstract—Reverberation Time (T_{60}) and Direct-to-Reverberant Ratio (DRR) are important parameters which together can characterize sound captured by microphones in non-anechoic rooms. These parameters are important in speech processing applications such as speech recognition and dereverberation. The values of T_{60} and DRR can be estimated directly from the Acoustic Impulse Response (AIR) of the room. In practice, the AIR is not normally available, in which case these parameters must be estimated blindly from the observed speech in the microphone signal. The Acoustic Characterization of Environments (ACE) Challenge aimed to determine the state-of-the-art in blind acoustic parameter estimation and also to stimulate research in this area. A summary of the ACE Challenge, and the corpus used in the challenge is presented together with an analysis of the results. Existing algorithms were submitted alongside novel contributions, the comparative results for which are presented in this paper. The challenge showed that T_{60} estimation is a mature field where analytical approaches dominate whilst DRR estimation is a less mature field where machine learning approaches are currently more successful.

Index Terms—Acoustic signal processing, parameter estimation, reverberation

I. INTRODUCTION

ACOUSTIC PARAMETERS can provide important information about the acoustic environment, and the speech quality and intelligibility observed in that environment. The performance of speech processing algorithms is widely seen now to be improved through awareness of relevant acoustic parameters, for example, dereverberation performance [1]–[4], and speech recognition performance [5]–[10].

Of all the available parameters that characterize room acoustics, the Acoustic Characterization of Environments (ACE) Challenge focused on Reverberation Time (T_{60}) and Direct-to-Reverberant Ratio (DRR).

The T_{60} is defined as the time taken for a sound in the diffuse sound field to decay by 60 dB after an abrupt cessation of the source [11], whilst the DRR is the energy ratio of the sound arriving at the observation point directly from the source, to the sound arriving at the observation point after being reflected from one or more surfaces [12].

The T_{60} and DRR can be estimated from a measured Acoustic Impulse Response (AIR) using existing methods [12], [13]. However, in many practical situations, the AIR is not available and so the T_{60} and DRR must be estimated non-intrusively from the observed reverberant speech. In this context, non-intrusively means that no prior information about the speech or the AIR is available.

Increasingly, additive noise is a problem for speech processing applications [14]. Gaubitch et al. [15] showed that three state-of-the-art T_{60} estimators [16]–[18] were significantly biased in the presence of additive noise. Since that review, further developments have been made in non-intrusive T_{60} estimation in noise such as [19]–[23]. Further, there have been advances in DRR estimation [24]–[27] which have not yet been the subject of a comparative evaluation. In addition, devices with multiple microphones are now commonplace with typical mobile telephone voice processing components supporting up to the three microphones [28]. Therefore spatial information is available in practical applications, which is useful in determining DRR [26], [27]. There is therefore a need to assess the state-of-the-art in non-intrusive T_{60} and DRR estimation in single and multiple microphone scenarios with additive noise.

A. Accurately simulating additive noise

In order to develop and test acoustic parameter estimation algorithms robust to noise such as [19], it is necessary to have available test signals from either real or simulated noisy reverberant environments. It is usually impractical to record comprehensive and representative speech in real acoustic environments since each speech utterance needs to be recorded in each acoustic environment which may involve many combinations of utterances, rooms, microphone positions, noise types, and noise levels. It is then restrictive in terms of creating new combinations of reverberant speech. For practical purposes, an assumption is therefore made that the AIR represents a Linear Time Invariant (LTI) system, which is a useful approximation for situations where the room conditions are static over the period of time necessary to make AIR measurements. Under this assumption it is possible to construct representative noisy reverberant speech by convolving anechoic speech recordings with measured AIRs, and mixing with noise.

An existing technique for introducing noise is to use pre-recorded monaural noise recordings for the noise source such as [29], and adding it to a speech signal that has been convolved with an AIR. How noise arrives at the microphone however is dependent on the characteristics of the room, and there exists a different AIR for every point from which noise emanates. The noise types of interest in testing of T_{60} and DRR estimation algorithms are numerous. In this study, babble J. Eaton, A. H. Moore, and P. A. Naylor are with Imperial College London. Nikolay D. Gaubitch is with TU Delft and Pindrop Security Manuscript received December 18, 2015; revised May 2, 2016.
noise, ambient noise, and fan noise have been considered. Babble noise is a common test for speech processing algorithms, representing a realistic scenario of speech in a crowded room, and is applicable to many practical applications including mobile communications and hearing aids. Ambient noise represents the background encountered in everyday situations usually in buildings. Fan noise represents situations where there is close proximity to a computer fan or air conditioning system, for example.

Accurately simulating different types of noise where the noise emanates from many or all parts of a room requires the capture or simulation of a very large number of AIRs. Simulating realistic room noise is generally only practical when adding perfectly diffuse noise sources. It is impractical, therefore, for adding noise emanating from many sources in the same room such as babble noise and non-stationary ambient noise since many individual measurements would need to be made for each source.

Of the corpora used in recent $T_{60}$ and DRR estimation research [30]–[33], there are no associated noise recordings. In a related study, the Reverb Challenge, the SimData corpus [34] incorporates ambient noise recorded under the same conditions as the AIRs. A further recent corpus, Distant-speech Interaction for Robust Home Applications (DIRHA) [35] incorporates recordings of domestic noises from rooms within a single apartment. Whilst ambient noise is provided in two of the corpora above, neither of these studies provides babble noise and fan noise recorded under the same conditions as the AIR measurements.

B. ACE Challenge and corpus

In order to stimulate research and to determine the state-of-the-art in non-intrusive acoustic parameter estimation in realistic noise, inspired by [15], the Acoustic Characterization of Environments (ACE) Challenge was devised involving the collection of a new corpus of multi-channel noisy reverberant speech. The ACE corpus comprises spontaneous anechoic speech, measured multiple microphone AIRs, and multiple microphone ambient, fan and live babble noise recorded under the same conditions as the AIRs. The ACE Challenge is one of several research challenges coordinated by the IEEE Audio and Acoustic Signal Processing (AASP) Technical Committee.

The contribution of this paper is to describe the ACE Corpus, to describe the process of the ACE Challenge and how the corpus was used in the challenge, and to perform comparative testing of existing and novel algorithms on the corpus data. In addition, this paper provides several comparative evaluations and analyses of the results. An earlier overview of the ACE Corpus and Challenge was previously presented in [36].

The remainder of the paper is organised as follows: In Section II, the ACE Corpus is described, along with the methods for determining the ground truth $T_{60}$ and DRR. In Section III, the ACE Challenge is described. In Section IV, a summary of the results of the ACE Challenge are provided, and in Section V, conclusions are drawn.

II. ACE Corpus

The ACE Corpus\textsuperscript{1} is a new database of anechoic speech, multiple microphone AIRs, and multiple microphone noise for providing realistic multi-channel noisy reverberant speech utterances for the development and evaluation of speech processing algorithms. Included are AIRs for 7 different rooms with 5 different microphone array configurations in 2 different microphone positions per room in order to provide a wide range of $T_{60}$ and DRRs. Importantly, it also includes babble, ambient and fan noise recorded under same conditions as the measured AIRs. The $T_{60}$ and DRR were obtained from the measured AIRs for each room, microphone position, and microphone, in both Fullband (FB), giving a single value for all frequencies up to the Nyquist rate, and ISO preferred frequency bands [37], giving a single value for each frequency band. These ground-truth parameter values are provided with the corpus. In addition, a set of anechoic recordings of spontaneous speech is provided. Realistic noisy reverberant speech signals can be constructed by combining the various elements of anechoic speech, AIRs and associated noise using software supplied with the corpus. These signals can span a wide range of $T_{60}$, DRR and Signal-to-Noise Ratio (SNR), and, whilst the corpus has been designed specifically to test acoustic parameter estimation algorithms, it is also highly relevant to many other tasks.

A. Acoustic model

A noisy reverberant speech signal, $y_m(n)$, with discrete-time index, $n$, captured in a room by a microphone $m$, is characterised by an AIR, $h_m$, of length $L_h$, of the acoustic channel between the source and microphone, where $h_m = [h_{m,0}, h_{m,1}, \ldots, h_{m,L_h-1}]^T$. It is assumed that the AIR is time invariant. The noisy reverberant signal incorporating the convolution of the speech with the AIR can be written as

$$y_m(n) = h_m^T \cdot x(n) + v_m(n), \quad (1)$$

where $x(n) = [x(n), x(n-1), \ldots, x(n - L_h + 1)]^T$ is the speech signal vector, and $v_m(n)$ is additive noise at the $m$th microphone. To obtain a noisy reverberant speech signal, $y_m(n)$, at the test SNR, $\xi$, a clean speech signal, $x(n)$, is convolved with the respective AIR, $h_m$, as in (1), and then mixed with noise as

$$y_m(n) = h_m^T \cdot x(n) + \frac{v_m(n)\sqrt{\xi_s}}{\sqrt{\xi}} \quad (2)$$

where $\xi_s$ is power ratio of the unprocessed speech and noise signals given by

$$\xi_s = \frac{A_s}{E[|v_m(n)|^2]} \quad (3)$$

and where $A_s$ is the active speech level of the reverberant speech utterance without noise estimated using ITU-T P.56 Method B [38], and $E[\cdot]$ is the expectation operator.

\textsuperscript{1}Available www.ace-challenge.org
B. Anechoic speech recordings

Two sets of speech signals were recorded. In Set 1, four male talkers were recorded and in Set 2, five female and five male talkers were recorded. The utterances comprise the subject speaking their favourite colour, the town where they live, a description of where they work, a description of how they travel to work, and counting 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 in a single sitting per talker per set. A B&K 4190-L-001 measurement microphone connected to a B&K NEXUS 2690 conditioning amplifier was used, and digitized using an RME FireFace 800 audio interface, using a sample rate of 48 kHz and 24-bit precision. The speech signals were then normalised to give approximately equal loudness measured in Loudness Units Full-Scale (LUFS), and then manually segmented into utterances which vary in length according to utterance type and speaker. Following the EBU recommendation R 128 [39], given a target loudness of −23 LUFS, the normalization ensured all speech signals were within ±1 LUFS of the target.

C. Rooms

Table I lists the dimensions for the seven different offices, meeting and teaching rooms within Imperial College London that were used to produce the corpus. Also included for each room is the volume, estimated mean T60 across all microphones and AIRs, and the range of estimated DRR for the two microphone positions, Position 1 and Position 2. The room characteristics are as follows:

Office 1: a carpeted office containing a table, desk and four chairs;
Office 2: a carpeted office containing a table, desk, 6 chairs and a bookcase;
Meeting Room 1: a carpeted meeting room containing a meeting table and 14 chairs;
Meeting Room 2: a carpeted meeting room containing approximately 30 chairs and 6 tables;
Lecture Room 1: a hard-floored lecture room containing approximately 20 tables and 60 chairs;
Lecture Room 2: a hard floored lecture room containing approximately 35 tables and 100 chairs;
Building Lobby: a large irregular-shaped hard-floored room with coupled spaces including a café, stairwell and staircase (shown in Fig. 1). Measurements in Table I correspond to the corner area where the recordings were made whereas the total volume of the lobby is many times larger. There was a high level of both stationary and non-stationary naturally occurring ambient noise including electrically operated doors, lifts with announcements, and building users walking by the recording environment.

D. Microphone positions, configurations, source and seating positions

The source position and the seating position of the occupants remained the same for all recordings within each room, whilst there were two separate positions per room of all the microphones, Position 1 and Position 2. Table II shows the microphone configurations used for all AIR and noise recordings in the corpus. All recordings were made sample-synchronously across all channels using a sample rate of 48 kHz and 24-bit precision.

The 3-element Mobile, 5-element Crucif and 8-element Lin8Ch arrays were recorded using two RME OctaMic preamps with their balanced outputs connected to the balanced inputs of two clock-synchronized RME FireFace 800s connected to a MacBook Pro. The recording software employed was Audacity. The 32-element Eigenmike recordings were made using the Eigenmike Microphone Interface Box (EMIB) clock-synchronized to the FireFace 800s connected to a second MacBook Pro. The 2-element Chromebook, recordings were made using arecord in little endian format so as to directly record the signals from the two integrated microphone, but not however clock-synchronised to the other audio interfaces.

In each room AIR measurements and noise recordings were made with the rooms occupied with the talkers for the babble noise. For most rooms unoccupied AIR measurements were also made. Further details of the recording procedure are provided in [40]. Figure 1 shows the recording equipment in place in the Building Lobby ready to commence recording.

E. Noise recordings

Three different noise types (ambient, fan and babble) were recorded in each room for each microphone position. To generate the babble noise, four to seven people sitting in the recording environment were asked to speak continuously for the duration of the noise recording. Talkers were provided with a list of phrases from TIMIT [41], or could bring their own material. In a few cases talkers read from scientific papers. To maintain as constant an acoustic environment as possible, talkers remained in their seated positions for the duration of the recording session of all noises and all AIR measurements. The ambient noise was recorded with occupants present but remaining silent. One or two fans in the room were used to create the fan noise, taking care to avoid creating wind noise.

For all of the rooms and microphone positions except Office 2, AIRs for the rooms without participants were captured. For both the fan and babble noises, there is also ambient noise in
the reverberant tail in the diffuse sound field, and the noise exponential decay of the reverberant signal as a single mode, the reverberant signal which includes three components: the centre frequency 

Thus the centre frequency measurement in T

obtained nominally the same for each room. 

channels and positions for each room, and minimum and maximum source-microphone distances for each room the mean values for T

below 30 ms (10 000 samples) once the level fell below −70 dB to remove noise and artefacts. The AIRs were analysed for FB and Subband (SB) T

respectively. Table I shows for each room the mean values for T

variances for each room were found to be less than 0.004 s² validating the assumption that T

is nominally the same for each room.

\( T_{60} \) estimates from the AIR

The method of Karjalainen et al. [13] was used to determine the ground truth T

in FB and in frequency bands using the ISO preferred centre frequencies [37]. Thus the centre frequency of band 1 is at 25.1189 Hz, and band 26 is at 7.943 kHz. This method applies a non-linear optimization to a model for the reverberant signal which includes three components: the exponential decay of the reverberant signal as a single mode, the reverberant tail in the diffuse sound field, and the noise floor. This was found to be more reliable under all conditions than either the ISO-3382 T

which tended to over-estimate the reverberation time in the rooms. This algorithm is suitable for where the decay has three sections such as is the case for the ACE Challenge, but otherwise the algorithm may not converge to the expected value of T

The filter bank used for the frequency-dependent T

Estimation was a 0.5-octave 8th order Butterworth design, designed using the Matlab fdesign.octave function.

\[ DRR_m = 10 \log_{10} \left( \sum_{n=0}^{n_0} h_{m,n}^2 + \sum_{n=n_0+1}^{n_\infty} h_{m,n}^2 \right) \]

where the direct path signal arrives at sample \( n_d \), and \( n_0 = 120 \) samples at a sample rate of 48 kHz, or 2.5 ms, which represents an additional path difference of 0.85 m at 340 m s⁻¹. In order to identify the sample pertaining to the direct path, \( n_d \), an equalization filter was applied to the AIR which compensated for the frequency response of the source loudspeaker and produced coefficients in the AIR representing distinct reflections. The location of the direct path, \( n_d \), was then determined by finding the maximum value of the AIR after equalization. The AIR without equalization was then used in (4). Frequency-dependent DRR estimates were made using the same method as for the frequency-dependent \( T_{60} \) estimates described in Section II-G. It is necessary to include the energy before the arrival of the direct path because of the influence of the sampling process.

<table>
<thead>
<tr>
<th>Name</th>
<th>Elements</th>
<th>Type</th>
<th>Alignment</th>
<th>Spacing</th>
<th>Element type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromebook</td>
<td>2</td>
<td>Chromebook Pixel</td>
<td>Horizontal</td>
<td>Laptop screen bezel, 02 mm</td>
<td>MEMS digital omni.</td>
</tr>
<tr>
<td>Mobile</td>
<td>3</td>
<td>Perspex former</td>
<td>Right-angled triangle</td>
<td>Base: 45 mm, side 100 mm</td>
<td>DPA4060 omni.</td>
</tr>
<tr>
<td>Crucif</td>
<td>5</td>
<td>Perspex former</td>
<td>4-arm with central mic.</td>
<td>250 mm centre-to-arm</td>
<td>DPA4060 omni.</td>
</tr>
<tr>
<td>LinLibCh</td>
<td>8</td>
<td>Perspex former</td>
<td>Linear</td>
<td>60 mm centre-to-centre</td>
<td>DPA4060 omni.</td>
</tr>
<tr>
<td>EM32</td>
<td>32</td>
<td>Eigenmic</td>
<td>Spherical</td>
<td>Distributed on 84 mm rigid sphere baffle</td>
<td>14 mm electret omni.</td>
</tr>
</tbody>
</table>

**TABLE I**

<table>
<thead>
<tr>
<th>Room properties</th>
<th>Microphone position 1</th>
<th>Microphone position 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>L (m)</td>
<td>W (m)</td>
</tr>
<tr>
<td>Office 1</td>
<td>3.22</td>
<td>4.83</td>
</tr>
<tr>
<td>Office 2</td>
<td>3.22</td>
<td>4.83</td>
</tr>
<tr>
<td>Meeting Room 1</td>
<td>6.61</td>
<td>5.11</td>
</tr>
<tr>
<td>Meeting Room 2</td>
<td>10.3</td>
<td>9.07</td>
</tr>
<tr>
<td>Lecture Room 1</td>
<td>6.93</td>
<td>9.73</td>
</tr>
<tr>
<td>Lecture Room 2</td>
<td>13.6</td>
<td>9.29</td>
</tr>
<tr>
<td>Building Lobby</td>
<td>4.47</td>
<td>5.15</td>
</tr>
</tbody>
</table>

**TABLE II**

**ACE CORPUS MICROPHONE CONFIGURATIONS**
III. THE ACE CHALLENGE COMPARATIVE EVALUATION

A. ACE Challenge Overview

The ACE Challenge comparative evaluation was devised to determine the state-of-the-art in non-intrusive acoustic parameter estimation - specifically \( T_{60} \) and DRR - and to stimulate research in this field. Central to the challenge was the noisy reverberant speech of the ACE Corpus described in Section II.

The following non-intrusive estimation tasks were set:

1) Single microphone \( T_{60} \) and DRR estimation;
2) Multi-microphone \( T_{60} \) and DRR estimation;
3) Single microphone \( T_{60} \) and DRR estimation in \( \frac{1}{3} \)-octave SBs;
4) Multi-microphone \( T_{60} \) and DRR estimation in \( \frac{1}{3} \)-octave SBs.

Participants were expected to use the Development (Dev) dataset to tune their algorithms for \( T_{60} \) and DRR estimation, and then test their algorithms on the fully blind Evaluation (Eval) dataset using the provided software tools. The organizers then decoded and analyzed the submitted results and returned them to the participants.

Participants were encouraged to enter multiple types of algorithms, using their preferred microphone configurations. They could then select the most promising algorithms to report in their paper submissions.

The ACE Challenge attracted participation from nine research teams from eight different countries around the world. Participants and their algorithms are listed in Table III in order of appearance of their algorithms in the results tables.

<table>
<thead>
<tr>
<th>Participant</th>
<th>Algorithms submitted (see results tables)</th>
<th>( T_{60} )</th>
<th>DRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Federal University of Rio de Janeiro (UFRJ)</td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Friedrich-Alexander-Universität (FAU), RWTH Aachen University</td>
<td>B, C, D, E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Imperial College London</td>
<td>F, O</td>
<td>k, l, m, n, o</td>
<td></td>
</tr>
<tr>
<td>Fraunhofer IDMT, Carl von Ossietzky University, Oldenburg</td>
<td>H</td>
<td>j</td>
<td></td>
</tr>
<tr>
<td>INRS-EMT, University of Quebec,</td>
<td>I, J, K, L, M, N, O, P, Q</td>
<td>p, q, r, s, t, u, v, w, x, y</td>
<td></td>
</tr>
<tr>
<td>Nuance Communications Inc.</td>
<td>R, S, T</td>
<td>i, g, h</td>
<td></td>
</tr>
<tr>
<td>Microsoft Research</td>
<td>U, V</td>
<td></td>
<td></td>
</tr>
<tr>
<td>University of Auckland, NTT Corporation</td>
<td>a, b, c, d, e</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Australian National University (ANU)</td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

B. ACE Challenge datasets

Dev and Eval datasets of noisy reverberant speech files were constructed from the ACE Corpus described in Section II. The purpose of the Dev dataset was to allow participants of the ACE Challenge to review the performance of their algorithms on typical ACE data, and to perform any tuning before commencing the challenge. The purpose of the Eval dataset was to provide fully blind noisy reverberant speech upon which to base the challenge.

Each noisy reverberant speech file was constructed from anechoic speech convolved with measured AIsRs obtained from a given room, combined with a random selection of the noise (ambient, fan or babble) recorded in the same room conditions, with the same occupants, and with the same source and microphone configuration using the method described in Section II-A.

The Dev dataset comprised noisy reverberant speech files from 2 rooms with 2 microphone positions each, 4 male talkers, 2 utterances each, babble, fan, and ambient noise at 0 dB, 10 dB, and 20 dB SNR for all microphone configurations. There were 288 files per microphone configuration. Ground truth \( T_{60} \) and DRR information for the Dev set was provided to participants for each microphone element for each microphone position in both FB and at all ISO preferred frequency bands [37] up to the Nyquist frequency.

The Eval dataset comprised noisy reverberant speech files from 5 rooms different to those provided in the Dev dataset, with 2 microphone positions each, 5 male and 5 female talkers, 5 utterances each, babble, fan and ambient noise in low, \(-1\) dB, medium, 12 dB, and high 18 dB SNR scenarios generated using the same method as for the Dev dataset. There were 4500 files per microphone configuration. The noisy reverberant speech files were numbered in a pseudorandom permutation, which was different for each microphone configuration, to reduce the possibility of training on the Eval dataset. Both Dev and Eval datasets were downsampled prior to distribution to a sample rate of 16 kHz and converted to 16-bit precision.

In both Dev and Eval datasets, a further single-channel microphone configuration was included. In the Dev dataset channel 1 of the Lin8Ch array was used. For the Eval dataset channel 1 of the Crucif array was used.

C. ACE Challenge software

Software was provided to participants for each phase to assist in generating results in the required format. The Dev software allowed participants to test their algorithms on the Dev dataset and analyse the results by SNR, by \( T_{60} \), and by DRR, from a database of ground truth values. The Eval software allowed participants to test their algorithms on the Eval dataset to produce data files suitable for submission to the ACE Challenge, and also to analyse the decoded results when returned. The software also recorded Real-Time Factor (RTF) for each test, the Direction-of-Arrival (DoA) azimuth and elevation if estimated, and the SNR if estimated for subsequent analysis.

D. Taxonomy of algorithms submitted

There were three main classes of algorithms submitted to the ACE Challenge:

1) Analytical with or without Bias Compensation (ABC);
2) Single Feature with Mapping (SFM);
3) Machine Learning with Multiple Features (MLMF).

The ABC approaches derive the estimate for the acoustic parameter directly from the signal without requiring any prior information. Bias compensation may be performed in order to
account for noise or specific aspects of the source material. An example of this is the maximum likelihood method [17] which directly produces the $T_{60}$ estimate.

The SFM approaches estimate a parameter from a signal that is correlated with the acoustic parameter to be estimated, and then apply a mapping function to give the acoustic parameter estimate. An example of this is the Spectral Decay Distributions (SDD) method [16] which determines Negative-Side Variance (NSV) from STFT bins and then applies a mapping to obtain the $T_{60}$.

The MLMF approaches typically use many features of the source material to train a machine learning algorithm such as a neural network which then estimates the acoustic parameter from the features of a test signal. An example of this is the Non-Intrusive Room Acoustics (NIRA) [47] algorithm.

There were no hybrid approaches submitted to the ACE Challenge although several participants applied noise reduction to the source signals before performing parameter estimation.

IV. RESULTS AND DISCUSSION

In this Section, a synoptic presentation of the comparative results is given. Unabridged results can be found in the ACE Technical Report [40]. Separate performance comparisons are provided for the FB tasks given in Section III-A in Figs. 2 and 3. These are presented as box plots where there is a box shown for each algorithm. For each box, the central notch is the median, the edges of the box are the 25th and 75th percentiles, and the whiskers extend to the most extreme data points not considered outliers. Outliers are plotted individually. The algorithms are identified on each box plot by a single character which corresponds to the character in the summary numerical results presented in Tables IV and V. Boxes are colour-coded according to algorithm class: ABC: yellow; MLMF: cyan; and SFM: green. The Pearson correlation coefficient, $\rho$, between the estimates and the ground truth for each algorithm is plotted as a blue cross in the same column as the algorithm. The results are sorted by participant by correlation coefficient across all noises and SNRs in FB.

For the $T_{60}$ FB task, the last three algorithms (W, X, Y) are those compared in Gaubitch et al. [15], and are included as baselines to enable the progress made in non-intrusive $T_{60}$ estimation since the earlier evaluation to be assessed. Similarly, for the DRR FB task, Jeub et al. [24] is included as a baseline algorithm since this was a freely available estimator prior to the ACE Challenge. The columns in Tables IV and V are as follows:

1) Ref., the identifier for each algorithm used on the horizontal axis of the preceding figure;
2) Algorithm, the name used by the respective ACE Challenge participant to refer to their algorithm;
3) Class, the class of algorithm according to Section III-D;
4) Mic. Config., the microphone configuration of the Evaluation dataset used to test the algorithm;
5) Bias, the mean error in the estimation results;
6) MSE, the mean squared error in the estimation results;
7) $\rho$, the Pearson correlation coefficient between the estimation results and the ground truth;
8) RTF, the real-time factor, the total computation time divided by the total duration of all processed speech files.

All implementations were in Matlab except for those marked with a $\dagger$ which used Matlab for feature extraction and C++ for the machine learning-based mapping, and those marked with a $\dagger$ which were implemented entirely in C++.

By considering the bias, MSE, and $\rho$ together, it is possible to determine how well the estimator performs in these tests. For example, an estimator with a low bias and MSE might give an estimate close to the median for every speech file. By examining $\rho$, it is possible to distinguish between such an algorithm, which will have $\rho \to 0$, and an algorithm which is more accurately estimating the parameter concerned, which will have $\rho \to 1$. The RTF is useful for indicating whether the algorithm has the potential to be used in practical applications that are constrained in computational complexity such as hearing aids and mobile devices.

A. Results by correlation coefficient for $T_{60}$ estimation

The overall results for FB $T_{60}$ estimation are shown in Fig. 2 and Table IV. There were no multi-channel approaches submitted that exploited the spatial information within the AIRs. The $T_{60}$ is defined in the diffuse field where $\text{DRR} < 0 \text{dB}$, and several algorithms rely on this assumption. However, the median $\text{DRR}$ was approximately 5dB, so the task was challenging. In spite of this the best performing algorithms achieved a correlation coefficient of $\rho \approx 0.8$.

Around half the algorithms outperformed the baselines from Gaubitch et al. [15], with the best performing showing a significant improvement which suggests that non-intrusive $T_{60}$ estimation is a mature field with estimators performing well. The most accurate algorithms were SFM or ABC approaches. Features based on decay rates dominated the most accurate approaches as expected since decay rates are a defining characteristic of a reverberant decay tail.

1) Fullband single microphone $T_{60}$ estimation: The overall results for FB $T_{60}$ estimation are shown in Fig. 2 and Table IV.

![Fig. 2. FB $T_{60}$ estimation error in all noises for all SNRs. Boxes are colour-coded according to algorithm class: ABC: yellow; MLMF: cyan; SFM: green](image)
The best performing algorithms achieved a correlation coefficient of ρ ≈ 0.8. The most accurate algorithms were SFM or ABC approaches. Features based on decay rates dominated the most accurate approaches as expected since decay rates are a defining characteristic of a reverberant decay tail. Around half the algorithms outperform the baselines from Gaubitch et al. [15], with the best performing showing a significant improvement which suggests that non-intrusive $T_{60}$ estimation is a mature field with estimators performing well.

The algorithm with the highest ρ was algorithm A [48]. This SFM approach transforms the signal into the log magnitude Short Time Fourier Transform (STFT) domain. Consecutive frames of speech in frequency bands are then used to determine decays in the signal using a frame-based Energy Decay Function (EDF) in frequency bands. Decreasing gradients are identified and the median value of all gradients in each frequency band is determined. The $T_{60}$ estimate is obtained from a trained first order mapping function using the median value of all medians in all frequency bands. To provide robustness to noise, the algorithm assumes that there exists only noise in the first 500 ms of each noisy reverberant speech utterance, and this noise power is therefore used to estimate the SNR. One of two noise reduction algorithms is selected based on the estimated SNR, and then used to pre-process the signal before transforming to the STFT domain.

The algorithm with the next highest ρ was algorithm B [49]. This ABC algorithm transforms the signal into the STFT domain, and after denoising using a Minimum Mean Squared Error (MMSE) noise estimator, and transforming back into the time domain, pre-selects frames containing sound decays. The $T_{60}$ estimate for each frame is then calculated using a log-likelihood function assuming the signal is represented by a random variable with a Gaussian Probability Density Function (PDF). The frame-based estimate is then smoothed and averaged over all frames of the utterance to give the final value.

2) Fullband multiple microphone $T_{60}$ estimation: There were no multi-channel approaches submitted that exploited the spatial information within the AIs. However, the multi-channel SFM algorithms K to M, and O to Q used averaging to obtain better estimates. This family of algorithms use variants of the SRMR metric based on modulation domain information to estimate $T_{60}$ and DRR [52], [53]. The SRMR is obtained by first applying a 23-channel gammatone filterbank to the time domain signal. A Hilbert transform captures the envelope of the output of each filter, and thus the temporal dynamics information. These signals are then transformed into the STFT domain to obtain an 8 band modulation spectrum segmented into frames. The SRMR is then obtained by comparing the energy in the different bands of the modulation spectrum. The averaged per-channel SRMR metric and its variants over all the channels prior to mapping achieves an increase in ρ over the single microphone algorithm for the Mobile and Chromebook microphone configurations. This does not however yield an improvement with the 5-channel Cruciform therefore the averaging technique may not be successful in all conditions.

3) Frequency-dependent single microphone $T_{60}$ estimation: The only algorithm submitted for this task was algorithm D [49]. This divides the signal into frequency subbands using an octave filter bank. A model for frequency-dependent $T_{60}$ is used to compensate for estimation bias and thus give a more accurate estimate. In order to compare results with FB algorithms in the ACE Challenge, the mean of the frequency dependent $T_{60}$ over the range 400 to 1250 Hz was used as recommended in ISO 3382 [45]. Under these conditions, this method performs less well than the FB algorithm upon which it is based, achieving the fourth highest ρ of all the $T_{60}$ estimation algorithms. For further details on the algorithm performance by frequency band, refer to [40].

---

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Algorithm</th>
<th>Class</th>
<th>Mic. Config.</th>
<th>Bias</th>
<th>MSE</th>
<th>DR</th>
<th>SRMR</th>
<th>NSRMR</th>
<th>ABC</th>
<th>FB</th>
<th>MSE</th>
<th>DR</th>
<th>SRMR</th>
<th>NSRMR</th>
<th>ABC</th>
<th>FB</th>
<th>MSE</th>
<th>DR</th>
<th>SRMR</th>
<th>NSRMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>QA Reverb [48]</td>
<td>SFM</td>
<td>Single</td>
<td>-0.068</td>
<td>0.0648</td>
<td>0.778</td>
<td>0.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>Octave sf-based FB RTE [49]</td>
<td>ABC</td>
<td>Single</td>
<td>-0.104</td>
<td>0.0733</td>
<td>0.738</td>
<td>0.939</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>DCT-based FB RTE [49]</td>
<td>ABC</td>
<td>Single</td>
<td>-0.104</td>
<td>0.0766</td>
<td>0.711</td>
<td>1.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>Model-based SB RTE [49]</td>
<td>ABC</td>
<td>Single</td>
<td>-0.0196</td>
<td>0.0981</td>
<td>0.661</td>
<td>0.451</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>Baseline algorithm for FB RTE [49]</td>
<td>ABC</td>
<td>Single</td>
<td>-0.0432</td>
<td>0.14</td>
<td>0.387</td>
<td>0.0424</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>SDDSA-G retrained [50]</td>
<td>SFM</td>
<td>Single</td>
<td>0.0167</td>
<td>0.0927</td>
<td>0.608</td>
<td>0.0152</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>SDDSA-BC/19</td>
<td>SFM</td>
<td>Single</td>
<td>-0.0323</td>
<td>0.0803</td>
<td>0.6</td>
<td>0.0164</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>Multi-layer perception [51]</td>
<td>MLMF</td>
<td>Single</td>
<td>-0.0967</td>
<td>0.104</td>
<td>0.48</td>
<td>0.0578</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>Per acoust. band SRMR Section 2.5, 52</td>
<td>SFM</td>
<td>Single</td>
<td>-0.114</td>
<td>0.109</td>
<td>0.48</td>
<td>0.578</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>NSMR Section 2.4, 52, 53</td>
<td>SFM</td>
<td>Single</td>
<td>-0.0646</td>
<td>0.119</td>
<td>0.261</td>
<td>0.571</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>NSMR Section 2.4, 52, 53</td>
<td>SFM</td>
<td>Chromebook</td>
<td>0.012</td>
<td>0.116</td>
<td>0.291</td>
<td>1.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>NSMR Section 2.4, 52, 53</td>
<td>SFM</td>
<td>Mobile</td>
<td>-0.0504</td>
<td>0.0858</td>
<td>0.281</td>
<td>1.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>NSMR Section 2.4, 52, 53</td>
<td>SFM</td>
<td>Crucif</td>
<td>-0.0516</td>
<td>0.107</td>
<td>0.236</td>
<td>2.62</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>SRMR Section 2.3, 52</td>
<td>SFM</td>
<td>Single</td>
<td>-0.16</td>
<td>0.144</td>
<td>0.22</td>
<td>0.457</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>SRMR Section 2.3, 52</td>
<td>SFM</td>
<td>Chromebook</td>
<td>-0.105</td>
<td>0.132</td>
<td>0.221</td>
<td>0.829</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>SRMR Section 2.3, 52</td>
<td>SFM</td>
<td>Mobile</td>
<td>-0.133</td>
<td>0.12</td>
<td>0.228</td>
<td>1.26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>SRMR Section 2.3, 52</td>
<td>SFM</td>
<td>Crucif</td>
<td>-0.153</td>
<td>0.128</td>
<td>0.225</td>
<td>2.99</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>NIHav3 [47]</td>
<td>MLMF</td>
<td>Single</td>
<td>-0.192</td>
<td>0.151</td>
<td>0.302</td>
<td>0.899</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>NIHav1 [47]</td>
<td>MLMF</td>
<td>Single</td>
<td>-0.184</td>
<td>0.151</td>
<td>0.258</td>
<td>0.899</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>NIHav2 [47]</td>
<td>MLMF</td>
<td>Single</td>
<td>-0.179</td>
<td>0.198</td>
<td>0.199</td>
<td>0.907</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U</td>
<td>Blur kernel [54]</td>
<td>SFM</td>
<td>Single</td>
<td>-0.173</td>
<td>0.15</td>
<td>0.279</td>
<td>8.46</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>Blur kernel with sliding window [55]</td>
<td>SFM</td>
<td>Single</td>
<td>-0.0055</td>
<td>0.139</td>
<td>0.12</td>
<td>0.421</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>Temporal dynamics [19]</td>
<td>SFM</td>
<td>Single</td>
<td>-0.0304</td>
<td>0.211</td>
<td>0.209</td>
<td>0.362</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>Improved blind RTE [17]</td>
<td>ABC</td>
<td>Single</td>
<td>-0.0635</td>
<td>0.165</td>
<td>0.166</td>
<td>0.0259</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>SDD [16]</td>
<td>SFM</td>
<td>Single</td>
<td>0.463</td>
<td>305</td>
<td>0.00158</td>
<td>0.0221</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4) Frequency-dependent multiple microphone Reverberation Time estimation: There were no entries for this task.

B. Results by correlation coefficient for DRR estimation

Overall results for FB DRR estimation are shown in Fig. 3 and Table V. Values of $\rho \geq 0.6$ were achieved, with the best performing algorithm exploiting spatial information in an ABC approach. The algorithms in general have large biases and MSEs. The MLMF approaches are more successful than for the $T_{60}$ task. This suggests that the modelling and estimation of DRR is less well understood, and that there may be better features for DRR yet to be exploited by ABC techniques, and that therefore this field of research is not as mature as $T_{60}$ estimation.

1) Fullband single microphone DRR estimation: Algorithm f [47] was the single microphone algorithm with the highest $\rho$. This MLMF approach uses a Voice Activity Detector (VAD) to select frames containing speech, and then computes 134 features including those typically used in Automatic Speech Recognition (ASR) schemes, in addition to others including modulation domain features, and spectrum features from a scattering transformation applied to the signal. The training corpus comprised 40% noisy reverberant speech from corpora other than the ACE Corpus. The inclusion of this additional training data increased $\rho$ by around 0.05.

2) Frequency dependent single microphone DRR estimation: There were no entries for this task.

3) Fullband multiple microphone DRR estimation: The algorithm with the highest $\rho$ in this task was algorithm a [56]. This ABC approach compares the PSD in the STFT domain of two different beamformers, and by accounting for the respective gain of each beamformer determines the DRR. Noise robustness is achieved using a VAD to select frames most likely to contain speech. Calibration with the ACE Corpus was achieved using an offset to account for the difference in the assumptions in calculating the DRR for the ACE Corpus compared to the assumptions used by the algorithm related to (4).

C. Analysis of results by additional corpus parameters

In this Section, the performance of $T_{60}$ and DRR estimation algorithms is analysed by a parameter in addition to $T_{60}$ or DRR correlation coefficient. The parameters are $T_{60}$ or DRR range, SNR, talker gender, and utterance length. Single and multiple microphone algorithms are compared alongside each other.

1) Fullband $T_{60}$ estimation: As can be seen in Fig. 4, most single microphone $T_{60}$ estimators underestimated longer $T_{60}$s, and overestimated shorter $T_{60}$s. Estimating $T_{60}$ is more difficult for long $T_{60}$s using decay rates as a feature because the gradients are very shallow and a small change in gradient leads to a large change in the $T_{60}$ estimate. Illustrative examples of this are the SFM algorithms F and G. Underestimation in some cases was a consequence of the Dev set containing a smaller range of $T_{60}$s than the Eval set. Illustrative examples of this are the MLMF algorithms R and S.

As can be seen in Fig. 5, noise affected all algorithms. In general noise causes overestimation since $T_{60}$ estimation algorithms have difficulty distinguishing between noise and the noise-like tail of reverberation. The blur kernel algorithms U and V, whilst showing promising results did not perform well in the levels of noise used in the challenge. Algorithm Y, the SDD algorithm reviewed in [15] had one of the lowest $\rho$,
TABLE V

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Algorithm Class</th>
<th>Mic. Config.</th>
<th>Bias</th>
<th>MSE</th>
<th>$\rho$</th>
<th>RTF</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>PSD est. in beamspace, bias comp. [56]</td>
<td>ABC Mobile</td>
<td>1.07</td>
<td>8.14</td>
<td>0.577</td>
<td>0.757</td>
</tr>
<tr>
<td>b</td>
<td>PSD est. in beamspace (Row) [56]</td>
<td>ABC Mobile</td>
<td>5.9</td>
<td>41.8</td>
<td>0.577</td>
<td>3.17</td>
</tr>
<tr>
<td>c</td>
<td>PSD est. in beamspace v2 [56]</td>
<td>ABC Mobile</td>
<td>5.7</td>
<td>43</td>
<td>0.41</td>
<td>0.844</td>
</tr>
<tr>
<td>d</td>
<td>PSD est. by twin BF [26]</td>
<td>ABC Mobile</td>
<td>5.71</td>
<td>44.9</td>
<td>0.362</td>
<td>0.899</td>
</tr>
<tr>
<td>e</td>
<td>Spatial Covariance in matrix mode [25]</td>
<td>ABC Mobile</td>
<td>5.37</td>
<td>61.2</td>
<td>0.244</td>
<td>0.627</td>
</tr>
<tr>
<td>f</td>
<td>NIRAv2 [47]</td>
<td>MLMF Single</td>
<td>-1.85</td>
<td>14.8</td>
<td>0.558</td>
<td>0.899</td>
</tr>
<tr>
<td>g</td>
<td>NIRAv3 [47]</td>
<td>MLMF Single</td>
<td>-1.62</td>
<td>14.7</td>
<td>0.515</td>
<td>0.899</td>
</tr>
<tr>
<td>h</td>
<td>NIRAv1 [47]</td>
<td>MLMF Single</td>
<td>-1.64</td>
<td>15</td>
<td>0.507</td>
<td>0.899</td>
</tr>
<tr>
<td>i</td>
<td>Particle velocity [57]</td>
<td>ABC EM32</td>
<td>-2.38</td>
<td>10.4</td>
<td>0.449</td>
<td>0.134</td>
</tr>
<tr>
<td>j</td>
<td>Multi-layer perceptron [51]</td>
<td>ABC Chromebook</td>
<td>-5.71</td>
<td>44.9</td>
<td>0.362</td>
<td>0.627</td>
</tr>
<tr>
<td>k</td>
<td>DENBE no noise reduction [27]</td>
<td>ABC Chromebook</td>
<td>-4.25</td>
<td>34.1</td>
<td>0.558</td>
<td>0.899</td>
</tr>
<tr>
<td>l</td>
<td>DENBE spectral subtraction [58]</td>
<td>ABC Chromebook</td>
<td>-4.01</td>
<td>32.8</td>
<td>0.303</td>
<td>0.775</td>
</tr>
<tr>
<td>m</td>
<td>DENBE spec. sub. Gerkmann [27]</td>
<td>ABC Chromebook</td>
<td>-4.01</td>
<td>32.8</td>
<td>0.303</td>
<td>0.0449</td>
</tr>
<tr>
<td>n</td>
<td>DENBE filtered subbands [58]</td>
<td>ABC Chromebook</td>
<td>-3.71</td>
<td>20.6</td>
<td>0.259</td>
<td>0.829</td>
</tr>
<tr>
<td>o</td>
<td>DENBE FFT derived subbands [58]</td>
<td>ABC Chromebook</td>
<td>-5.1</td>
<td>34.3</td>
<td>0.269</td>
<td>1.04</td>
</tr>
<tr>
<td>p</td>
<td>Normalized Overall SRMR (NSRMR) Section 2.2. [52]</td>
<td>SFM Chromebook</td>
<td>-4.47</td>
<td>32.2</td>
<td>0.116</td>
<td>1.26</td>
</tr>
<tr>
<td>q</td>
<td>Overall SRMR (OSRMR) Section 2.2. [52]</td>
<td>SFM Chromebook</td>
<td>-4.32</td>
<td>32.2</td>
<td>0.148</td>
<td>1.58</td>
</tr>
<tr>
<td>r</td>
<td>OSRMR Section 2.2. [52]</td>
<td>SFM Mobile</td>
<td>-3.28</td>
<td>22.2</td>
<td>0.116</td>
<td>1.26</td>
</tr>
<tr>
<td>s</td>
<td>GSMR Section 2.2. [52]</td>
<td>SFM Mobile</td>
<td>-4.05</td>
<td>31</td>
<td>0.0814</td>
<td>2.62</td>
</tr>
<tr>
<td>t</td>
<td>GSMR Section 2.2. [52]</td>
<td>SFM Crunch</td>
<td>-2.88</td>
<td>22.3</td>
<td>0.0816</td>
<td>2.09</td>
</tr>
<tr>
<td>u</td>
<td>GSMR Section 2.2. [52]</td>
<td>SFM Single</td>
<td>-4.16</td>
<td>33.9</td>
<td>0.00841</td>
<td>0.54</td>
</tr>
<tr>
<td>v</td>
<td>GSMR Section 2.2. [52]</td>
<td>SFM Single</td>
<td>-4.24</td>
<td>34.6</td>
<td>0.00815</td>
<td>0.446</td>
</tr>
<tr>
<td>x</td>
<td>Per acoust. band SRMR Section 2.5. [52]</td>
<td>SFM Single</td>
<td>-0.9</td>
<td>22.8</td>
<td>0.000192</td>
<td>0.578</td>
</tr>
<tr>
<td>y</td>
<td>Temporal dynamics [59]</td>
<td>SFM Single</td>
<td>-11.4</td>
<td>147</td>
<td>0.0815</td>
<td>0.0822</td>
</tr>
<tr>
<td>z</td>
<td>QA Reverb [48]</td>
<td>SFM Single</td>
<td>-2.51</td>
<td>23.6</td>
<td>0.0376</td>
<td>0.391</td>
</tr>
<tr>
<td>w</td>
<td>Blind est. of coherent-to-diffuse energy ratio [24]</td>
<td>ABC Chromebook</td>
<td>-12.1</td>
<td>162</td>
<td>0.305</td>
<td>0.019</td>
</tr>
</tbody>
</table>

which is a satisfying observation in terms of scientific progress since this earlier study.

As can be seen in Fig. 6, most algorithms performed better on male speech, with the deviation increasing as correlation coefficient decreases. The Dev set comprised only male talkers so an algorithm tuned on this database is more likely to show a larger estimation error with female talkers.

Figure 7 shows correlation by utterance length. Most algorithms performed better with longer utterances. Whilst algorithms A to E were the most accurate for longer utterances achieving $\rho \approx 0.9$, they were adversely affected by utterances shorter than 5 s. These algorithms both use the same method for selecting frames containing relevant decay information and these results suggest that this method is significantly less effective for utterances shorter than 5 s.

2) Fullband DRR estimation: Figures 8 and 9 show the results for DRR estimation for the Single and Mobile microphone configurations respectively for different ranges of DRR. These are shown separately. The values of $\rho$ are very low, particularly for DRR < 5 dB. Only algorithms a to c achieved $\rho \approx 0.5$ for DRR < 5 dB. The low values of $\rho$ in this test suggest that most algorithms did not work reliably as estimators under the test conditions, and that DRR estimation is not a solved problem.

As can be seen in Fig. 10, most algorithms were affected by noise. In the case of the modulation domain-based SFM approaches p to y, the addition of noise improved the correlation in noise for the multiple microphone algorithms.

As shown in Fig. 11, talker gender had little impact on
Fig. 7. FB $T_{60}$ estimation error in all noises and all SNRs for utterance lengths <5 s, utterance lengths < 15 s, and utterance lengths ≥ 15 s.

Fig. 8. Single microphone FB DRR estimation error in all noises and all SNRs for DRR < 2 dB, 2 ≤ DRR < 5 dB, and DRR ≥ 5 dB.

Fig. 9. Mobile (3-channel) FB DRR estimation error in all noises and all SNRs for a) DRR < 2 dB b) 2 ≤ DRR < 5 dB and c) DRR ≥ 5 dB.

Fig. 10. FB DRR estimation error in all noises at 18 dB SNR, 12 dB SNR, and -1 dB SNR.

Fig. 11. FB DRR estimation error in all noises and all SNRs for female and male talkers.

Fig. 12 shows that short utterance lengths affected some of the algorithms adversely. Algorithms f to o showed a significant dependency on utterance length. These algorithms all use different features and approaches to estimate the DRR. This suggests that the information being used is more sparsely distributed amongst the noisy reverberant speech, and that longer utterances are required to obtain sufficient information to make an accurate estimate.

V. Conclusion

The ACE Challenge aimed to determine the state-of-the-art in blind $T_{60}$ and DRR estimation by collecting and sharing a new corpus containing anechoic speech, a large and diverse set of measured AIRs, and condition-matched noise. It also aimed most algorithms. However, the MLMF algorithms g, h, and j, were trained exclusively on the Dev dataset comprising only male talkers, and as expected performed less well with female speech. Algorithm f used additional training data from other corpora that were not exclusively male and was more robust to female speech.

Figure 12 shows that short utterance lengths affected some of the algorithms adversely. Algorithms f to o showed a significant dependency on utterance length. These algorithms all use different features and approaches to estimate the DRR. This suggests that the information being used is more sparsely distributed amongst the noisy reverberant speech, and that longer utterances are required to obtain sufficient information to make an accurate estimate.

V. Conclusion

The ACE Challenge aimed to determine the state-of-the-art in blind $T_{60}$ and DRR estimation by collecting and sharing a new corpus containing anechoic speech, a large and diverse set of measured AIRs, and condition-matched noise. It also aimed
to stimulate new research. A wide variety of algorithms were submitted and, as a result of the challenge, several participants developed novel algorithms, and improved upon and extended existing algorithms. It is concluded therefore that the aims of the challenge have been met.

The results show that non-intrusive $T_{60}$ estimation is a mature field, where the performance of non-intrusive estimators has significantly improved since the study in [15]. Non-intrusive DRR estimation however is a significantly less mature field. Joint-estimation algorithms did not outperform algorithms focused on a single task, and their performance overall leads to the conclusion that the main distinguishing features of $T_{60}$ and DRR are different. For non-intrusive $T_{60}$ estimation, the current best-performing algorithm can estimate $T_{60}$ to within an RMS error of about 250 ms and a $\rho \approx 0.8$ for typical operating scenarios of $-1$ to 18 dB SNR. For non-intrusive DRR estimation, the best performing algorithm can estimate DRR to within an RMS error of about 3 dB and a $\rho \approx 0.6$ for typical operating scenarios of $-1$ to 18 dB SNR.

In this corpus the LTI assumption has been made when measuring the AIRs, recording the noises, and simulating noisy reverberant speech. In a real situation there will be additional head movements, room temperature variations, and air currents which challenge the LTI assumption. This is the subject for future investigations.

REFERENCES


