Robust Speech Recognition in Reverberant Environments by Using an Optimal Synthetic Room Impulse Response Model

Jindong Liu^{a,*}, Guang-Zhong Yang^a

^a*The Hamlyn Centre, Imperial College London, 180 Queens Gate, London, United Kingdom*

Abstract

This paper presents a practical technique for Automatic speech recognition (ASR) in multiple reverberant environments based on multi-model selection. Multiple ASR models are trained with artificial synthetic room impulse responses (IRs), i.e. *simulated* room IRs, with different reverberation time (T_{60}^{Model} s) and tested on *real* room IRs with varying T_{60}^{Room} s. To apply our method, the biggest challenge is to choose a proper artificial room IR model for training ASR models. In this paper, a generalised statistical IR model with attenuated reverberation after an early reflection period, named attenuated IR model, has been adopted based on three time-domain statistical IR models. Its optimal values of the reverberation-attenuation factor and the early reflection period on the recognition rate have been searched and determined. Extensive testing has been performed over four real room IR sets (63 IRs in total) with variant T_{60}^{Room} s and speaker microphone distances (SMDs). The optimised attenuated IR model had the best performance in terms of recognition rate over others. Specific considerations of the practical use of the method have been taken into account including: i) the maximal training step of T_{60}^{Model} in order to get the minimal number of models with acceptable performance; ii) the impact of selection errors on the ASR caused by the estimation error of T_{60}^{Room} ; and iii) the performance over SMD and direct-to-reverberation energy Ratio (DRR). It is shown that recognition rates of over

*Corresponding author

Email addresses: j.liu@imperial.ac.uk, +44 20 7594 5472 (Jindong Liu), gzy@imperial.ac.uk (Guang-Zhong Yang)

80~90% are achieved in most cases. One important advantage of the method is that T_{60}^{Room} can be estimated either from reverberant sound directly [1, 2, 3] or from an IR measured from any point of the room as it remains constant in the same room [4], thus it is particularly suited to mobile applications. Compared to many classical dereverberation methods, the proposed method is more suited to ASR tasks in multiple reverberant environments, such as human-robot interaction.

Keywords:

*speech recognition, reverberant environment, artificial synthetic room
impulse response*

1. Introduction

Although automatic speech recognition has been explored extensively [5], it still has many inherent limitations when applied to complex human-computer interactions (e.g. social mobile robots and intelligent home appliances) in real-life environments with unavoidable artefacts, due to reverberation. For instance, a natural human-robot communication interface is essential for a healthcare mobile robot and it often occurs in reverberant rooms with larger speaker-microphone distance (>1.5 m). Fig. 1 demonstrates a typical office setup in which room-specific reverberations can affect the ASR greatly. For practical applications, speech recognition needs to be robust to these changes as the robot roams around in different rooms. The performance of existing models trained with anechoic speech signals can deteriorate when the person talking to the robot is located a few metres away [6]. Thus far, many algorithms for ASR in reverberant rooms have been developed with a focus mainly on spectrum enhancement, feature enhancement, hidden Markov model (HMM) adaptation and reverberant modeling during speech recognition [7]. Existing research uses multiple channel input [8][9] to deal with background noise or simultaneous speech. However, many of them are not designed for real-time processing tasks [10], thus limiting their practical use. In this paper, we aim to apply multi-model selection to perform real-time speech recognition in reverberant environments.

The human auditory system can adapt to environmental changes such as ambient noise, reverberation differences between rooms, and even specific locations within a room. This adaptation capability has inspired recent research into new computational auditory models for complex human-computer

Figure 1: Concept illustration of robot speech recognition in multiple reverberant environments.

interactions.

One of the most well-known phenomena related to the reverberant sound perception of a human is the precedence effect [11]. Here, two spatially-separated sound stimuli, with a small time delay between them, are perceived as if from a single phantom spatial position if the inter-stimulus delay (ISD) is very small (less than 1 to 5 ms). The precedence effect was originally explained by a bottom-up process in the auditory system where those neurons responding to early arriving sound subsequently suppress the responses to later sound. However, recent studies have shown that humans also use top-down processes to deal with reverberation, in which a visit to a familiar auditory scene does not need any build-up time to adapt to the reverberation condition [12]. It is suggested that humans may apply top-down memory from the cortex to change the behaviour of the auditory system and that the human brain may maintain a multi-model speech recognition system according to variant reverberation conditions.

Based on the above biological evidence, we suggest an ASR to maintain multiple speech recognition models for various *auditory scenes* and to train these models from reverberant sound directly. Thus far, there are already several papers based on such an idea. However, most of them have not considered the more realistic situations where reverberation is location specific. For example, Matassoni [13] and Sehr [14] performed experiments using reverberant speech obtained from real room IRs for both training and testing. However, IR varies at different positions of the room, and thus it is difficult to measure the room IR during the ASR in different positions. This makes

it difficult to select the optimal model during run-time. In our previous work [15], this problem is overcome by associating pre-measured room IR with visual features of the room in the training stage and then recalling the corresponding speech recognition model based on visual scene recognition. Couvreur [16][17] trained models using reverberant speech from simulated room IRs and the evaluation was also performed on simulated data. However, the results on real room reverberant sound were problematic and the mismatching involved introduced significant recognition error. One common problem with these methods is that the training and testing are within the *same* setting. These approaches are impractical for many complex human-computer interactions, such as mobile robots, for two reasons: i) exact room IRs are difficult to obtain online for unknown geometries; and ii) IRs are dependent on the measurement position even in the same room. A practical scheme should ideally be trained with a set of generic, possibly simulated, IRs and should be generalisable to unseen environments.

The basic procedure can be divided into two steps: i) to train in advance a number of ASR models with statistical room IRs of varying reverberation time T_{60}^{Model} s, and ii) to choose the model that matches to the real room IR according to T_{60}^{Room} for practical ASR. One important advantage of the method is that T_{60}^{Room} can either be estimated from reverberant sound directly [1, 2, 3], or from an IR measured from any point in the room as it is constant throughout one room [4]. Although the procedure itself is not new, compared to classical dereverberation methods, it is more practical for real-time human-robot interaction tasks as no preprocessing is required for reverberant speech signals.

The main contribution of the paper is to find the optimal parameters for a generic IR model for training ASR models and evaluate its practical performance. A generalised statistical IR model with attenuated reverberation after an early reflection period, named attenuated IR model, has been optimised according to recognition rate. In the attenuated IR model, reverberation after an early reflection period is attenuated by a reverberant-attenuation factor, i.e. the direct and early reflection components are amplified. It is biologically inspired by the fact that the early reflections arriving before some 40 or 50 ms are beneficial for human speech intelligibility [18]. This fact has also been verified by the ASR when applying a clean speech model to reverberation-suppressed reverberant speech [19]. In this paper, four important practical questions have been addressed: i) What are the optimal values of the reverberant-attenuation factor and the early reflection

period for ASR? ii) How many models should be trained in advance? iii) What happens if the T_{60}^{Room} is different from T_{60}^{Model} ? (Or what is the impact of selection error on ASR due to unconfident T_{60}^{Room} estimation?) iv) Does the performance decay with increasing the speaker-microphone distance (SMD)?

The remainder of this paper is organised as follows: Section 2 describes the system model and a dedicated speech recognition training procedure. Section 3 shows the procedure for looking for the optimal values of the reverberant-attenuation factor and the early reflection period. In Section 4, experimental results are presented to demonstrate the feasibility and performance of the entire system proposed, and a detailed performance comparison of four statistical IR models to a classical speech recognition system is provided. This is then followed by discussion and conclusions in Section 5.

2. Multi-model Selection Speech Recognition

The proposed system consists of two main stages: training and testing, as shown in Figure 2. In the training stage, simulated room IRs with different T_{60}^{Model} s are generated. Then an anechoic utterance training data set is convolved with these IRs to generate a reverberant utterance set. Speech features from the reverberant data are then extracted to train a speech recognition model for each reverberation time. During the test phase, the model first estimates the room reverberation time T_{60}^{Room} from reverberant sound, i.e. the time it takes for the energy of the sound to reduce by 60 dB after the sound source is switched off. It then chooses the closest T_{60} -dependent speech recognition model for recognition. In this section, we illustrate the technical details involved in each stage.

2.1. Statistic Room Impulse Responses

A real room IR typically has three main components in time domain [4]: direct sound, early reflection and reverberation. The early reflection is right after the direct sound and before the reverberation. The reverberation generally starts from about 50 ms after the direct sound. There are several methods for generating artificial acoustic impulse responses. The source-image method [20], for example, considers the room geometry and the positioning of the source and the microphone. Such methods are impractical for our cases as the position of the microphone and the room geometry are changing during the operation. For this reason, we have not included the source-image method for this paper. On the other hand, time-domain statistical room IR

Figure 2: A schematic diagram of the main processing steps involved in the T_{60} -dependent speech recognition system.

models are based on random noise attenuated with a decaying exponential whose rate depends on the assumed reverberation time. The reverberation time is generic and does not necessarily depend on a specific room geometry. This makes these models suitable for the application scenarios considered in this paper. For the experiments considered, four time-domain statistical IR models have been taken into account. Three of them can be generalised by adding a reverberant-attenuation factor after an early reflection period, named attenuated IR model, for ASR. The remainder of this section will briefly introduce the four IR models and the attenuated IR model.

A familiar statistic IR model was proposed by Schroeder [21][22] in Equ. 1 which describes a room IR as a multiplication of a random sequence $b(k)$ and an exponentially decaying function $e^{-\alpha k/f_s}$.

$$h(k) = b(k)e^{-\alpha k/f_s}, k > 0 \quad (1)$$

where $b(k)$ is a white zero-mean Gaussian stationary noise. f_s is the sampling rate. α is the decay rate associated with the reverberation time T_{60} :

$$\alpha \triangleq \frac{3 \ln 10}{T_{60}}. \quad (2)$$

Schroeder's model was made for room acoustic measuring in which far field is the general assumption. Schroeder's model has no direct sound impulse or realistic early reflections. Houtgast [23] added a direct impulse by introducing a delta pulse into the discrete form of Schroeder's model shown in Equ. 3. However, the frequency dependence of IR has not been taken into account in both models. Recently, a frequency dependent IR model has been

proposed by Erkelens [24] shown in Equ. 4 in which amplitude A_j , decay constant δ_j , and phases ϕ_j , correspond to the modal frequencies ω_j of the room. A frequency dependent IR model is not included in the present study and refers to future work. We have included this model only for the purpose of comparison in this paper.

$$h(k) = \begin{cases} 1 & k = 1 \\ b(k)e^{-\alpha k/f_s} & k > 1 \end{cases} \quad (3)$$

$$h(k) = \begin{cases} 1 & k = 1 \\ \sum_{j=1}^J A_j e^{-\delta_j k} \cos(\omega_j k + \phi_j) & k > 1 \end{cases} \quad (4)$$

A generalised Schroeder's model was proposed by Habets [25] as in Equ. 5 by separately representing the early reflection and reverberation.

$$h(k) = \begin{cases} b_d(k)e^{-\alpha k/f_s} & 0 < k \leq T f_s \\ b_r(k)e^{-\alpha k/f_s} & k > T f_s \end{cases} \quad (5)$$

where T is the finishing time of the early reflection, typically 5 ms if only *very early reflections* are considered. The energy ratio between $b_d(k)$ and $b_r(k)$ is related to the Direct-to-Reverberation energy Ratio (DRR). Some researchers have shown that DRR can be blindly estimated from reverberant speech [26]. If a room IR is known, the DRR can be calculated from the following equation with $t_d = 5$ ms:

$$C_{t_d} = 10 \log_{10} \left(\sum_{k=1}^{t_d f_s} h^2(k) \right) / \sum_{k=t_d f_s+1}^{\infty} h^2(k) \quad (6)$$

However, it is not feasible to train multiple models with varying DRR and T_{60}^{Model} together because DRR depends on SMD. There will be too many models to train for all possible combinations of SMD and T_{60}^{Model} . Inspired by the biological finding that early reflection within about 50 ms is helpful to human speech intelligibility, a generalised statistical IR model with a reverberant-attenuation factor can be written as the following equation:

$$h(k, T, A) = \begin{cases} b(k)e^{-\alpha k/f_s} & 0 < k \leq T f_s \\ 10^{-\frac{A}{20dB}} b(k)e^{-\alpha k/f_s} & k > T f_s \end{cases} \quad (7)$$

where T is the early reflection period varying from 5 ms to 100 ms. A is a reverberant-attenuation factor to amplify the direct and early reflection

components. Their values will be fixed for all models once they are optimised in order to reduce the number of training models. This IR model is named as *Attenuated IR* in this paper. Note that the three previously mentioned IR models are the special case of this generalised attenuated model. For example, it will turn into Schroeder’s IR model in the case of $A = 0$ dB and it is the same as Habets’ IR model in the case of $10^{-\frac{A}{20dB}}b(k) = b_r(k)$ with the same T value.

Once a simulated room IR is obtained, the reverberant sound signal can thus be approximated by:

$$x_R(k, T, A) = \sum_{m=0}^{N-1} h(m, T, A)s(k - m) \quad (8)$$

Figure 3 shows a snapshot of the Energy Decay Curve (EDC) of the above four IRs together with two real room IRs as an example. EDC is calculated from the backward integration of the squared room impulse response, $h(t)$, according to [27]. It indicates that room IRs can be formulated into one and the same generic model as Equ. 7 by tuning the parameters with the early reflections time and late reverberation energy.

$$EDC(k) = 10 \log_{10} \left(\sum_{m=k}^{\infty} h^2(m) \middle/ \sum_{m=1}^{\infty} h^2(m) \right) \quad (9)$$

2.2. Model Training for Speech Recognition

Once the reverberant training data is prepared, it can be imported into any of the existing speech recognition training modules to obtain the T_{60} -dependent model. We used a general benchmark speech recogniser, HTK [30], as an example to test our system. Speech features for each training wave were extracted into standard 39-dimensional Mel frequency cepstral coefficients (MFCCs) [31], including 12 Mel-cepstral coefficients and the logarithmic frame energy, plus the corresponding delta and acceleration coefficients. Finally, all MFCC features were used to train word-level Hidden Markov Models (HMMs) with a left-to-right model topology. They assume no skips over states and have 32 Gaussian mixtures per state with diagonal covariance matrices [32].

In this paper, we adapted the Grid corpus [33] for clear training sound as in Figure 2. The Grid corpus consists of utterances of 34 human speakers

recorded in an anechoic room. Each utterance is a sentence with the syntax of

$$\langle command \rangle \langle colour \rangle \langle prep \rangle \langle letter \rangle \langle number \rangle \langle coda \rangle \quad (10)$$

Each speaker is assigned a set of words, following this syntax, to be read out loud. We randomly selected 500 sentences from the full set of 1000 sentences for each speaker for training the speech recogniser. The remaining 500 sentences were used for evaluation. These anechoic sentences were convolved with each of the simulated IRs, creating 17,000 training sentences for each room IR. For comparison, we also trained a speech recognition model on the anechoic sentences. The HTK model training code was adapted from the CHiME challenge [34].

2.3. Model Selection and Testing

In the testing stage, the T_{60}^{Room} estimation of the testing speech can be performed in two ways: either from a measurement of a room impulse response [27] or from the reverberant speech [1, 2, 3]. In our experiments, the reverberation time was estimated from the impulse responses following the ISO3382 standard [35].

A least-squares fitting process is then applied to the range between -5 and -35 dB. The reverberation time is finally computed from the slope of the line.

In practice, T_{60}^{Room} can be estimated from reverberant speech. For example, Löllmann [3] proposed a blind estimation method from reverberant speech. The algorithm reduced the computational cost by a downsampling operation and an efficient pre-selection of possible sound decays. It is also capable of estimating the T_{60}^{Room} in the presence of moderate background noise. Unoki [36] proposes a method of blindly estimating T_{60}^{Room} based on the modulation transfer function (MTF) from the reverberant signal. It was evaluated in real reverberant environments and shown that it could accurately estimate T_{60}^{Room} .

Having T_{60}^{Room} we then choose the speech recognition model trained on the T_{60}^{Model} that is the closest to T_{60}^{Room} . Suppose we have N T_{60} -dependent models, $M(T_{60}^{\text{Model}_i}), i = 1 \dots N$. The best matching model, $M(T_{60}^{\text{Model}_j})$, can be determined by:

$$j = \arg \min_i (|T_{60}^{\text{Model}_i} - T_{60}^{\text{Room}}|) \quad (11)$$

It is worth noting that the T_{60}^{Room} in the same room is relatively constant and there is no need to measure it for each testing position.

3. Optimal Parameters of Attenuated IR Model

As mentioned earlier, the attenuation factor A and the early reflection period T in Equ. 7 are fixed for varying T_{60}^{Model} s in order to limit the number of training models. Their values should be determined before subsequent experiments. To look for the optimal value according to the recognition rate of ASR, we tested the attenuated IR model with a range of values of A and T with the Aachen’s room impulse response (AIR) database [28]. The set of T ($N_T = 13$) is $T = [5, 10, 20, 30, 40, 50, 60, 80, 100, 200, 300, 500, 700]$ ms and the set of A ($N_A = 7$) is $A = [0, 5, 10, 15, 20, 25, Inf]$ dB. The reason for choosing $T > 100$ ms is to be consistent to the test in [19]. The case of $A = 0$ dB is equal to the Schroeder’s IR model. AIRs were recorded using a dummy head equipped with a binaural microphone, in realistic environments including an office, a meeting room, a lecture room, a stairway and a church hall. Recordings in an anechoic booth were collected as well as reference. The IRs were also recorded with varying speaker-microphone distances, ranging from one to ten metres. Such a range represents typical distances involved in human-robot interaction. Each room is labelled with a unique ID for the figures presented in this paper and the condition of each room and its corresponding T_{60}^{Room} are listed in Table 1. It is notable that our T_{60}^{Room} is not fully consistent with the given values in [28]. After correspondence with Marco Jaub, the author, we all agreed that our method of T_{60}^{Room} calculation is more accurate as it is compliant with ISO3382.

For each combination of T and A , 7 models with different T_{60}^{Model} s ($N_{T60} = 7$), i.e. $T_{60}^{\text{Model}} = [200, 400, 600, 800, 1000, 1500, 2000]$ ms, were trained so that 637 ($N_T * N_A * N_{T60} = 13 * 7 * 7 = 637$) models are trained in total. After the training, they were tested on the dataset of Aachen’s 23 IRs by matching T_{60}^{Model} with T_{60}^{Room} as shown in Section 2.3.

Figure 4 shows the results of average recognition rate, $RR(T, A)$ over T and A . The results were divided into two groups according to T_{60}^{Room} of AIRs: (i) for IRs with $T_{60}^{\text{Room}} < 1$ s and (ii) all IRs, i.e. $T_{60}^{\text{Room}} < 4$ s. The IRs in the first group are the most common case in our daily life. The results show that the peak values were roughly within a bounding box, i.e. $10 < T < 50$ ms and $10 < A < 15$ dB. Within this box, the average recognition rates were above 87% for the IRs in the first group and above 84% for those in the

second group. Outside the box, the result drop when $A < 10$ dB and decrease quickly with increasing A values if $T \leq 100$ ms. These results declined to the bottom when $A = Inf$ dB and $T \leq 100$ ms. The lowest value was at the point of $A = Inf$ dB and $T = 5$ ms. When $T > 100$ ms, the result hardly varied regardless of A .

In order to determine a combination of T and A for following experiments, four combinations of T and A were chosen from the highest recognition rates in the second group: RR(10, 15)=86.54%, RR(30, 10)=86.88%, RR(30, 15)=86.86% and RR(50, 10)=86.44%. Their results for each Aachen’s IR are plotted in Figure 5 and highlighted in 5-(a), (b) and (c) with either green or red colour. The performance of the model deteriorated with increasing T_{60}^{Room} especially when $T_{60}^{Room} > 0.5$ s. This is consistent with the findings in previous research [37]. The best overall result was obtained when A was between 5 dB to 15 dB. The results also show an interesting converging trend according to increasing T value. Over the four combinations, the results of RR(10, 15) and RR(30, 15) had the highest values in low T_{60}^{Room} band, i.e. $T_{60}^{Room} < 0.5$ s, and relatively lower values in other two T_{60}^{Room} bands ($0.5 < T_{60}^{Room} < 1$ s and $T_{60}^{Room} > 1$ s) especially from room 10 to 15. The results of RR(30, 10) and RR(50, 10) in Figure 5-(b) and (c) had a very similar performance. RR(50,10) has slightly better results in high T_{60}^{Room} band, such as room 20 and 21. Overall, we decided to choose the combination of $T = 50$ ms and $A = 10$ dB as the optimal parameters of the attenuated IR model.

4. Experimental Results

To evaluate the performance of these IR models and find the best one for complex human-computer interactions, we employed a Pioneer mobile robot equipped with a binaural microphone. The basic setup is shown in Fig. 1. Two external ear models were made by simulating the structure of the human ear and two cardioid microphones (Core Sound, NJ) were placed inside the ears, with 2.8 cm ear canals corresponding to the average canal length of adults [38]. A directional mono-speaker was used as a sound source. Note that only the left channel signal of the binaural microphone was processed. As the same robot had been used for binaural sound localisation, we keep the same setting for consistency.

Once the robot has estimated the T_{60}^{Room} of the current room, it uses the corresponding speech recogniser to recognise the sound recorded in the scene. We played the utterance from the Grid corpus through the speaker to test the

Index	Room ID	SMD (m)	T_{60}^{Room} (s)	DRR (dB)	Clarity Index (dB)
1	BT	1	0.08	8.59	24.91
2	BT	1.5	0.09	5.15	19.27
3	MT	1.45	0.32	6.30	17.36
4	MT	1.9	0.34	3.95	16.47
5	MT	1.7	0.34	5.42	17.16
6	MT	2.25	0.35	5.10	16.07
7	MT	2.8	0.35	5.34	15.73
8	OF	1	0.51	6.24	13.70
9	OF	3	0.55	-0.10	7.90
10	OF	2	0.55	2.11	9.40
11	LT	4	0.84	0.94	6.64
12	LT	2.25	0.84	3.65	8.35
13	LT	5.56	0.87	-0.97	4.86
14	LT	8.68	0.88	-5.15	4.85
15	LT	10.2	0.89	-0.87	5.68
16	LT	7.1	0.91	-1.91	4.15
17	SW	1	0.93	9.02	12.66
18	SW	2	1.11	4.02	8.53
19	SW	3	1.18	1.27	6.22
20	AL	1	2.80	6.98	11.55
21	AL	2	3.05	1.63	7.37
22	AL	3	3.23	0.34	5.39
23	AL	5	3.36	2.52	4.33

Table 1: The room ID and SMD of tested room IRs from Aachen Impulse Response (AIR) database [28]. The T_{60}^{Room} , DRR ($T_d=5$ ms) and Clarity Index ($T_d=50$ ms) are calculated. BT stands for booth, MT for meeting room, OF for office, LT for lecture hall, SW for stairway and AL for Aula Carolina hall. The DRR and Clarity Index are calculated using Equ. 6

recogniser. In order to evaluate our model in a controlled environment, we simulated the reverberant sound in the same way as with the training data, but using a different set of anechoic utterances and *real* measured room IRs. Four sets of room IRs, 63 in total, were convolved to generate reverberant speech for testing: Aachen Impulse Response database (23 IRs) with varying T_{60}^{Room} and SMD, Petrick’s IR with varying T_{60}^{Room} (10 IRs) [29], Petrick’s Smart Room IR (15 IRs) [39] with varying SMD, and our own Huxley IR (15 IRs) which was recorded using the robot. The T_{60}^{Room} of these IRs varied from 0.08 s to 3.36 s which is representative of a range of room geometries from small meeting rooms to large lecture halls. The SMD varies from 1 to 10 metres, which is the natural distance for human-robot communication.

4.1. Performance of Different Statistical IR models

In addition to the four statistical IR models mentioned above, two more reference models were chosen and a total of six speech recognition models were evaluated to compare their performances. They included: i) an ideal model (marked as “Ideal” in the following figures), in which the training data and testing data are both generated using the *same* room IR, similarly to the scheme in [13]; ii) Schroeder’s IR model (marked as “Schroeder IR”), in which the training data is generated from simulated IRs using Equ 1. Such a model was proposed in [17]; iii) attenuated IR model (marked as “AttenIR”), in which the training data was generated from simulated IRs using Equ. 7 with $A = 10$ dB and $T = 50$ ms; iv) the anechoic model that was trained using anechoic utterance only, marked as “Baseline”. (Such a training method is applied to most of the existing speech recognition algorithms); v) Houtgast’s IR Model (marked as “Houtgast IR”) in which Equ. 3 was applied to generate the training data; and vi) Frequency dependent IR model (marked as “fDep_IR”), in which the energy in each frequency band was calculated over all IRs and Equ. 4 was then applied to generate simulated IRs for the training data. Note that models i) and vi) are impractical for our application scenarios and they were only chosen as references.

Figure 6 shows the speech recognition results for reverberant sound using the six models. Recognition of only one of the two words was classified as unsuccessful. The training point of T_{60}^{Model} was every 200 ms from 200 to 1000 ms plus two long T_{60}^{Model} cases, 1500 ms and 2000 ms. The baseline model has the lowest recognition rate, with their best results being when $T_{60}^{\text{Room}} < 0.5$ s. In fact, most are below 50% when $T_{60}^{\text{Room}} > 0.5$ s. It is unsurprising that the “Ideal” model has the highest recognition rate over all T_{60}^{Room} s, as both the

training and testing used the IR of the same room. The Schroeder’s IR model provides recognition rates above 70% for most rooms of $T_{60}^{\text{Room}} < 1.5$ s but decays quickly under 50% above that. The attenuated IR model shows the best overall performance over all non-ideal models, which achieves above 80% recognition rate in most of the cases, which is very close to that of the ideal model. Houtgast’ IR model and the Frequency dependent IR model have a similar result over all rooms. Their performances are slightly worse than Schroeder’s model when $T_{60}^{\text{Room}} \leq 0.5$ s and almost overlap with Schroeder’s IR’s results when $0.5 < T_{60}^{\text{Room}} < 2.8$ s.

To assess the performance of the model in other conditions, we measured the IRs of 15 rooms in our building with the same robot used in [15]. These real room IRs were labelled as Huxley IRs. Figure 7 shows the dimension of all the rooms. These rooms include six different types: office, laboratory, corridor, meeting room, lecture room and kitchen. The head of the robot and its ear pinna are different from those in the Aachen IR database. The reverberant test sentences were generated using these room IRs and tested with all models. Apart from the Huxley IRs, Petrick’s real room IRs used in [29] were also tested. Figure 8 and Figure 9 show the results generated for these rooms. It is evident that the attenuated IR model remains the closest match to the “Ideal” model. The interesting fluctuation of the recognition rate over the SMD will be discussed in the following section.

4.2. Effect of T_{60}^{Room} Estimation Error

During testing, one of the important factors to consider is the estimation of T_{60}^{Room} from reverberant sound because a poor estimation of T_{60}^{Room} can cause a mismatch between the model and data. We are interested to see how robust the attenuated IR model is in the presence of a T_{60}^{Room} estimation error. The evaluation was performed by choosing a range of T_{60}^{Room} centred at the true T_{60}^{Room} value of testing data in order to simulate the estimation error, i.e. $T_{60}^{\text{Room}} = \tilde{T}_{60} \pm \Delta T_e$, where $\tilde{T}_{60}$ is the true T_{60}^{Room} value and ΔT_e is the error boundary. We choose ΔT_e within the range of $\Delta T_e = [0, 0.3]$ in steps of 0.05 s. Then we applied the proposed method at each ΔT_e and collected all possible recognition results. The results derived are shown in Figure 10 where $\tilde{T}_{60}$ is divided into three groups: less than 0.5 s, between 0.5 and 1 s, and larger than 1 s. From these figures, it can be seen that attenuated IR model is not sensitive to the estimation error in the last group. For the first and second groups, if we can control the $\Delta T_e < 0.2$ s then the mean recognition rate can be maintained above 90% and 85% respectively, and the

variance is within 5%. In the latest blind T_{60}^{Room} estimation method [3], the estimation error can be achieved within 0.1 s. Therefore, it is practical and feasible to control $\Delta T_e < 0.2\text{s}$.

4.3. Effect of T_{60}^{Model} Training Step Length

In terms of computational cost, most of the processing involved is during multi-model training, where the number of models is the same as the number of preselected T_{60}^{Model} s. Thus the way to discretise T_{60}^{Model} in a predefined range can greatly affect the computational cost of training. The evaluation therefore needs to assess whether a fine selection of T_{60}^{Model} can improve the recognition rate and how large a T_{60}^{Model} step can be between two neighbouring trained models. To study such effects, we trained more models with finer T_{60}^{Model} steps between 0.1 to 1.5 s and broader T_{60}^{Model} ranging up to 3.5 s. All models were organised into 5 groups according to the T_{60}^{Model} step (see Table 2 for the T_{60}^{Model} step of each group and the trained T_{60}^{Model} points). Note that groups 1 to 4 have inconsistent T_{60}^{Model} steps after $T_{60}^{\text{Model}} = 1.5$ s, as we only have four testing IRs in Aachen’s IR set which have T_{60}^{Room} greater than 1.5 s. Figure 11 shows the corresponding results for each group. It shows that the recognition rate is worst when the step is 500 ms. Overall, the groups with T_{60}^{Model} steps of 100, 200, 300 and 400 ms have very similar results. This indicates that a fine T_{60}^{Model} step is not necessary in training and a reasonable choice is a step of 400 ms or less. It is noticeable that room 1 (Booth) has decreasing performance when T_{60}^{Model} step increases, because it is an anechoic room. Although the recognition rate dropped to 90% for the 400 ms step, such type of room is rare for our cases. Further investigation was conducted into the relationship between the T_{60} difference, i.e. $\Delta T_{60} = T_{60}^{\text{Room}} - T_{60}^{\text{Model}}$, and the recognition rate. Figure 12 shows the corresponding results. When $T_{60}^{\text{Room}} < 1.2$ s (room 1 to 19), the maximum recognition rate occurs in the vicinity of 0, which is to be expected. In contrast, when $T_{60}^{\text{Room}} > 2.5$ s (room 21 to 23), the peak of the recognition rate is around $\Delta T_{60} = 1.5$ s, which indicates that the best model for this data is $T_{60}^{\text{Model}} = 1.4$ or $T_{60}^{\text{Model}} = 1.5$. In conclusion, it is sufficient to use 300 or 400 ms steps up to 1.5 s for modelling.

4.4. Effect of SMD and DRR

In general, two key parameters are often used to describe reverberation for room IRs: the reverberation time T_{60} and the direct-to-late reverberation ratio (DRR). DRR changes according to SMD, but the T_{60} is relatively constant for a given room. In our room impulse response model, we assumed

Table 2: Models to test T_{60}^{Model} step length.

Group No.	T_{60}^{Model} Step	Trained T_{60}^{Model} point (s)
1	100 ms	0.1: 0.1: 1.5, 2.0:0.5:3.5
2	200 ms	0.2: 0.2: 1.4, 1.5:0.5:3.5
3	300 ms	0.3: 0.3: 1.5, 2.0:0.5:3.5
4	400 ms	0.4: 0.4: 1.2, 1.5:0.5:3.5
5	500 ms	0.5: 0.5: 3.5

that T_{60} is the main factor for ASR. To investigate the SMD effect, we reorganised the attenuated IR model’s results (Figure 6) according to SMD. See Figure 13. It indicates that the results were roughly constant or slightly smeared according to SMD when T_{60}^{Room} is less than 1.5 s. SMD affected the result mainly in the large T_{60}^{Room} cases such as room AL. Figures 14 and 15 rearrange the recognition results in Figure 6 according to DRR and Clarity Index. They show that the results of attenuated IR model have the least variance compared to other IR models. The jumping points occurred when $T_{60}^{\text{Room}} > 1$ s such as the cases of AL_3m ($T_{60}^{\text{Room}} = 3.23$ s) and SW_2m ($T_{60}^{\text{Room}} = 1.11$ s).

Another test was performed on Petrick’s SmartRoom IR database in which IRs were collected in the same room, called SmartRoom, with varying SMDs. Figure 16 shows the results over all statistical IR models. It can be concluded that the attenuated IR model has the smallest variation effected by the SMD, although the recognition rate declined over all models when SMD increased. The attenuated IR model also has the best results over all the statistical models across all SMDs. It is evident that the DRR does not greatly affect the recognition rate of the statistical models trained with reverberant speech, but can have a significant impact on the anechoic model (Baseline). Given the fact that T_{60} in most typical rooms is less than 1.5 s, attenuated IR model is an acceptable practical solution for mobile ASR applications.

5. Discussion and Conclusions

In this paper, we have demonstrated an applicable technique for speech recognition in multiple reverberant environments. Inspired by the human ability to perceive sound in variable reverberant conditions, the proposed technique used a multi-model selection method to train multiple speech recog-

nisers characterised by the reverberation time T_{60} , and chose the T_{60} -matched recogniser to conduct speech recognition. A generalised Schroeder IR model or Habets' IR model, was revised for practical usage by using a reverberation-attenuation factor to amplify the direct and early reflection components before an early reflection time. The extensive test on real room IRs indicate that our proposed method can deal with reverberant sound effectively, with consistently high recognition rates when compared to traditional anechoic training methods and other IR models. The attenuated room IR is characterised by T_{60} , which is generic and does not depend on sampling position in a room. It enables the ASR to apply only one T_{60} -dependent speech recogniser anywhere in the room. One of the advantages of the technique is that real-time processing is guaranteed. The average processing time of 3 s speech is only 1 s on an Intel Core 2 Duo 2.13 GHz processor.

The reverberation-attenuation factor A and early reflection period T in the attenuated IR model mainly affect the performance of the model in ASR. In order to determine their optimal values, a range of A and T was tested and it was found that the optimal value is within a bounding box of $10 < T < 50$ ms and $10 < A < 15$ dB. The combination of $T = 50$ ms and $A = 10$ dB was a reasonable choice for most practical applications.

Regarding the effect of T_{60}^{Room} on estimation error, a number of scenarios were simulated. It was shown that attenuated IR model was not sensitive to the estimation error when $T_{60}^{\text{Room}} > 1.5$ s. When $T_{60}^{\text{Room}} < 0.5$ or $0.5 < T_{60}^{\text{Room}} < 1$ s, the mean recognition rate can be expected to be above 90% or 85% respectively if the estimation error is less than 0.2 s.

A potential drawback of our model is that multiple speech recognisers have to be trained in advance. In this paper, the number of recognisers was determined by the range of T_{60}^{Model} and the step size between adjacent recognisers. In order to reduce the number of recognisers, the effect of T_{60}^{Model} step size was studied. Results have shown that it is adequate to use 300 or 400 ms steps up to 1.5 s for training. This requires only 4 or 5 recognisers during training, which is reasonable for practical applications.

One of the assumptions in our model is that T_{60} is the main factor for artificial speech recognition and in this paper we have not taken into account the direct-to-late reverberation ratio (DRR) in room IRs. In order to justify this assumption, the effect of speaker distance on recognition rate was studied. The results showed that although DRR did affect all statistical IR models, the attenuated IR model had the least performance drop and kept above 80% in most of the cases. This is acceptable for practical applications.

In conclusion, by use of the attenuated IR model and multiple pre-trained speech recognisers, an ASR model can perform real-time speech recognition in variable reverberant environments by using the proposed method. The model has been evaluated on both public databases and our own to demonstrate its practical value for mobile human-robot interaction. In future research, the underlying reason for why the attenuated IR model performed better should be investigated. Nevertheless, we have already performed a test on the EDC as shown in Figure 3 which shows that the Attenuated IR model with $A = 10$ dB has a close match to the real room IR, although it may be too early to offer further conclusions. Word models of ASR were chosen in this paper as it is suitable for our applications, such as robots, to take simple commands from users. The difference of the effect of reverberant training on a phoneme based HMM from a word based HMM should be investigated in the next step.

Acknowledgment

We would like to thank Ning Ma and Heidi Christensen at the University of Sheffield for their suggestions on HTK's training code. Thanks to Nikolay Gaubitch for the suggestion on the reverberation time measurement from impulse response. The database and the code used in the paper can be shared upon request.

References

- [1] R. Takeda, K. Nakadai, T. Takahashi, K. Komatani, T. Ogata, H. G. Okuno, Automatic estimation of reverberation time with robot speech to improve ICA-based robot audition, in: Humanoid Robots, 2009. Humanoids 2009. 9th IEEE-RAS International Conference on, 2009, pp. 250–255.
- [2] T. H. Falk, W.-Y. Chan, Temporal dynamics for blind measurement of room acoustical parameters, IEEE Transactions on Instrumentation and Measurement 59 (4) (2010) 978–989.
- [3] H. Löllmann, E. Yilmaz, M. Jeub, P. Vary, An improved algorithm for blind reverberation time estimation, in: Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC), 2010, pp. 1–4.

- [4] H. Kuttruff, Room Acoustics, Fourth Edition, 4th Edition, Taylor & Francis, 2000.
- [5] J. Baker, L. Deng, J. Glass, S. Khudanpur, C. Lee, N. Morgan, D. O'Shaughnessy, Research developments and directions in speech recognition and understanding. Part 1, IEEE Signal Processing Magazine 26 (3) (2009) 75–80.
- [6] J. Baker, D. Li, S. Khudanpur, L. Chin-Hui, J. Glass, N. Morgan, D. O'Shaughnessy, Updated MINDS report on speech recognition and understanding, Part 2, Signal Processing Magazine, IEEE 26 (4) (2009) 78–85.
- [7] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, W. Kellermann, Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition, IEEE Signal Processing Magazine 29 (6) (2012) 114–126.
- [8] J. M. Valin, S. Yamamoto, J. Rouat, F. Michaud, K. Nakadai, H. G. Okuno, Robust recognition of simultaneous speech by a mobile robot, IEEE Transactions on Robotics 23 (4) (2007) 742–752.
- [9] C. T. Ishi, S. Matsuda, T. Kanda, T. Jitsuhiro, H. Ishiguro, S. Nakamura, N. Hagita, A robust speech recognition system for communication robots in noisy environments, IEEE Transactions on Robotics 24 (3) (2008) 759–763.
- [10] P. A. Naylor, N. D. Gaubitch (Eds.), Speech Dereverberation, Springer, 2010.
- [11] R. Litovsky, H. Colburn, W. Yost, S. Guzman, The precedence effect, Journal of the Acoustical Society of America 106 (4) (1999) 1633–1654.
- [12] T. Djelani, J. Blauert, A. Seestern, Modelling the direction-specific build-up of the precedence effect, in: Forum Acusticum, Sevilla, Spain, 2002.
- [13] L. Calmes, G. Lakemeyer, H. Wagner, Azimuthal sound localization using coincidence of timing across frequency on a robotic platform, The Journal of the Acoustical Society of America 121 (4) (2007) 2034–2048.

- [14] A. Sehr, C. Hofmann, R. Maas, W. Kellermann, A novel approach for matched reverberant training of HMMs using data pairs, in: 11th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2010, pp. 566–569.
- [15] J. Liu, E. Johns, G.-Z. Yang, A scene-associated training method for mobile robot speech recognition in multisource reverberated environments, in: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2011, pp. 542–549.
- [16] L. Couvreur, C. Couvreur, C. Ris, A corpus-based approach for robust ASR in reverberant environments, in: 1st Annual Conference of the International Speech Communication Association (INTERSPEECH), 2000, pp. 397–400.
- [17] L. Couvreur, C. Couvreur, Blind model selection for automatic speech recognition in reverberant environments, *The Journal of VLSI Signal Processing* 36 (2) (2004) 189–203.
- [18] R. Y. Litovsky, B. Rakerd, T. C. Yin, W. M. Hartmann, Psychophysical and physiological evidence for a precedence effect in the median sagittal plane, *Journal of neurophysiology* 77 (4) (1997) 2223–2226.
- [19] A. Sehr, E. A. Habets, R. Maas, W. Kellermann, Towards a better understanding of the effect of reverberation on speech recognition performance, in: *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2010.
- [20] J. B. Allen, D. A. Berkley, Image method for efficiently simulating small-room acoustics, *The Journal of the Acoustical Society of America* 65 (4) (1979) 943–950.
- [21] M. R. Shchroeder, Modulation transfer functions: definition and measurement, *Abstract in IEEE Transactions on Acoustics Speech and Signal Processing* 26 (1978) 180.
- [22] M. R. Schroeder, Modulation transfer functions: definition and measurement, *Acustica* 49 (3) (1981) 179–182.

- [23] T. Houtgast, H. Steeneken, R. Plomp, Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics, *Acustica* 46 (1) (1980) 60–72.
- [24] J. S. Erkelens, R. Heusdens, A statistical room impulse response model with frequency dependent reverberation time for single-microphone late reverberation suppression., in: 12th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2011, pp. 201–204.
- [25] E. Habets, Single- and multi-microphone speech dereverberation using spectral enhancement, Ph.D. Thesis, Technische Universiteit Eindhoven, The Netherlands (Jun. 2007).
- [26] M. Jeub, C. Nelke, C. Beaugeant, P. Vary, Blind estimation of the coherent-to-diffuse energy ratio from noisy speech signals, in: 19th European Signal Processing Conference (EUSIPCO), 2011, pp. 1347–1351.
- [27] M. R. Schroeder, Integrated-impulse method measuring sound decay without using impulses, *Journal of the Acoustical Society of America* 66 (2) (1979) 497–500.
- [28] M. Jeub, M. Schafer, P. Vary, A binaural room impulse response database for the evaluation of dereverberation algorithms, in: 16th International Conference on Digital Signal Processing, 2009, pp. 1–5.
- [29] R. Petrick, K. Lohde, M. Lorenz, R. Hoffmann, A new feature analysis method for robust asr in reverberant environments based on the harmonic structure of speech, in: 16th European Signal Processing Conference (EUSIPCO), 2008.
- [30] J. Odell, D. Ollason, P. Woodland, S. Young, J. Jansen, *The HTK book for HTK V2. 0*, Cambridge University Press, Cambridge, UK, 1995.
- [31] P. Mermelstein, Distance measures for speech recognition, psychological and instrumental, *Pattern Recognition and Artificial Intelligence* 116 (1976) 91–103.
- [32] M. Cooke, J. R. Hershey, S. J. Rennie, Monaural speech separation and recognition challenge, *Computer Speech & Language* 24 (1) (2010) 1–15.

- [33] M. Cooke, J. Barker, S. Cunningham, X. Shao, An audio-visual corpus for speech perception and automatic speech recognition, *Journal of the Acoustical Society of America* 120 (2006) 2421–2424.
- [34] H. Christensen, J. Barker, N. Ma, P. Green, The CHiME corpus: a resource and a challenge for computational hearing in multisource environments, in: 11th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2010, pp. 1918–1921.
- [35] ISO3382, Acoustics – Measurement of the reverberation time of rooms with reference to other acoustical parameters, International Organisation for Standardisation, Geneva, Switzerland, 2009.
- [36] M. Unoki, S. Hiramatsu, MTF-based method of blind estimation of reverberation time in room acoustics, in: 16th European Signal Processing Conference (EUSIPCO), 2008.
- [37] R. Petrick, T. Fehér, M. Unoki, R. Hoffmann, Methods for robust speech recognition in reverberant environments: A comparison, in: 11th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2010, pp. 582–585.
- [38] S. Mehrgardt, V. Mellert, Transformation characteristics of the external human ear, *The Journal of the Acoustical Society of America* 61 (6) (1977) 1567–1576.
- [39] R. Petrick, X. Lu, M. Unoki, M. Akagi, R. Hoffmann, Robust front end processing for speech recognition in reverberant environments: utilization of speech characteristics, in: 9th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2008, pp. 658–661.

Figure 3: A comparison of the energy decay curve (EDC) over all statistical IR models. T_{60}^{Model} was set to 0.8 s to match two real room IRs in which $T_{60}^{\text{Room}} = \{0.84, 0.86\}$ s and $\text{SMD} = \{2.25, 3\}$ m. The early reflection period T in Equ. 7 is fixed as 50 ms. LT_2.25m_0.84s was from Aachen's IR [28] and RicoIR_3m_0.86s from Petrick's IR [29].

Figure 4: The average recognition rate ($RR(T,A)$) over 23 Aachen's IRs according to the attenuation factor A and the early reflection period T . Top row shows $RR(T,A)$ for all IRs in which T_{60}^{Room} is less than 1 s. Bottom row shows the $RR(T,A)$ for all IRs in which T_{60}^{Room} is less than 4 s. In the left column, $RR(T,A)$ s are illustrated by heatmap. In the right column, the same results are plotted by 3D surface with logarithmic scale on T axis.

Figure 5: The recognition rate of each Aachen's IR according to the attenuation factor A and the early reflection period T . The results are arranged according to the T_{60}^{Room} of testing IRs. The room index corresponds to the same name column in Table 1. Note that $T_{60}^{\text{Room}} < 0.5$ s if room index is less than 8 and $T_{60}^{\text{Room}} > 1$ s if room index is larger than 18. Figures c, d, e, f share the same legend.

Figure 6: Recognition results of reverberant speech using six types of ASR models on Aachen's real room IR database. The room ID is arranged in ascending order of T_{60}^{Room} .

Figure 7: The floor plan for the part of the building where the experiments were conducted in our laboratory.

Figure 8: Recognition results of reverberant sound using our own IR database which was recorded in the building as shown in Figure 7. The rooms are ordered according to their T_{60}^{Room} s (the room IDs correspond to those shown in Figure 7). The SMD for all IRs are 2 m except for KC and MR in which SMD=1 m.

Figure 9: Recognition results of reverberant sound using Petrick's real room IR database [29]. The rooms are ordered according to their T_{60}^{Room} s. The room ID given by Petrick consists of the rounded T_{60}^{Room} value in microseconds, together with the SMD.

Figure 10: T_{60}^{Room} estimation error effect on recognition rate in different T_{60}^{Room} bands. The height of bars indicate the average, and the length of sticks indicate the variance of recognition rate.

Figure 11: The effect of T_{60}^{Model} step length during training, on the recognition rate. Only odd numbers of room index are plotted for clarity.

Figure 12: The relationship between ΔT_{60} and the recognition rate, where $\Delta T_{60} = T_{60}^{\text{Room}} - T_{60}^{\text{Model}}$. The legends indicate the room ID and the corresponding T_{60}^{Room} value. Only odd numbers of room index are plotted for clarity.

Figure 13: The speaker distance effect on recognition rate of Aachen's IR database. The T_{60} values in the legend are the average value of each room.

Figure 14: Recognition results of reverberant speech using six types of ASR models on Aachen's real room IR database. The room ID is arranged in ascending order of DRR in Table 1

Figure 15: Recognition results of reverberant speech using six types of ASR models on Aachen's real room IR database. The room ID is arranged in ascending order of Clarity Index in Table 1.

Figure 16: The speaker distance effect on recognition rate of Petrick's SMART room IR database. The average T_{60}^{Room} value is 0.83 s.