The Harming Part of Room Acoustics in Automatic Speech Recognition

Rico Petrick¹, Kevin Lohde², Matthias Wolff¹ and Ruediger Hoffmann¹

¹Dresden University of Technology, Laboratory of Acoustics and Speech Communication
²Dresden University of Applied Sciences, Department of Electrical Engineering

Abstract
Automatic speech recognition (ASR) systems used in real indoor scenarios suffer from different noise and reverberation conditions compared to the training conditions. This article describes a study which aims to find out what are the most harming parts of reverberation to speech recognition. Noise influences are left out. Therefore different real room impulse responses in different rooms and different speaker to microphone distances are measured and modified. The results of the recognition experiments with the related convoluted impulse responses clearly show the dependency of early and late as well as high and low frequency reflections. Conclusions concerning the design of a dereverberation method are made.

Index Terms: Room impulse response, reverberation, ASR

1. Introduction
Closing the gap of the behavior of ASR systems between laboratory and real acoustic conditions has employed a lot of research effort in the past. A common model to describe the disturbances of speech in a real acoustic environment is

\[ x(t) = (s * h)(t) + n(t) \] (1)

where \( x(t) \) is the signal at the input of the speech recognizer, \( n(t) \) is the sum of all noise signals which add at the microphone and \( h(t) \) is the room impulse response (RIR) between the speaker’s mouth and the microphone. Many investigations were made on how to deal with disturbing noises \( n(t) \). More recent is the research on how to deal with the influences of the surrounding room. Although a few approaches to compensate the room influences came up (a survey is found in [1] and as examples [2] and [3] are mentioned), speech recognition in a reverberant room is still a challenging task. To find a method of compensating the effects of reverberation this study investigates the influences of several components of the RIR dependent on the room and the speaker to microphone distance (SMD). Isolating the reverberation effects demands zero additive noise (\( n(t) = 0 \)) which is set throughout the experiments described here. This reduces (1) to a single convolution, but it is still a blind problem, since both \( s(t) \) and \( h(t) \) are unknown. The sampling frequency for all experiments is \( f_s = 16,000 \) Hz including the evaluation corpus, the ASR system and the RIRs.

2. Room Impulse Responses
2.1. Model of the room as an acoustic system
When observing the impulse response of a room different components of the RIR with different behavior can be supposed. Figure 1 describes the behavior of the room by 3 subsystems which can be described by their properties:

1 Direct sound: \( h_{d0} \) describes the sound wave traveling directly from the speaker’s mouth to the microphone without any reflections (it can be measured in an anechoic chamber).
   - It is represented as a single impulse in the RIR.
   - It has a delay which can be derived with the SMD and the sound velocity.
   - The amplitude is proportional to \( \frac{1}{r^2} \).
   - The frequency dependence of the direct sound can be neglected [4].

2 Discrete reflections: \( h_{dc} \) describes single reflections at a wall or at any other reflector. They rather appear in large rooms. In small rooms discrete reflections come together with diffuse reflections. Because of this, the later evaluations do not consider \( h_{dc} \).
   - They are represented as single impulses in the beginning phase of the RIR.
   - The behavior in delay and amplitude corresponds to the direct sound except of a frequency dependent attenuation at the reflection points according to the material and the geometrical structure of the surface.

3 Diffuse reflections - reverberation: Diffuse reflections described by \( h_e \) occur after the sound waves have passed multiple reflection points.
   - The reverberation phase in the RIR starts as a colored noise according to frequency dependent attenuation at reflection points.
   - Its mean energy degrades exponentially [5].
   - The degradation is frequency dependent (degrading faster for higher frequencies) for the following reasons:
     - Air sound attenuation [4]: (Frequency dependent) air attenuation becomes significant according to greater sound traveling distances after a number of diffuse reflections. (E.g. after 0.6 s (≈ 200 m) air attenuation becomes ≈ 4 dB at 1 kHz and ≈ 20 dB at 10 kHz.)
     - Frequency dependent reflections caused by different materials and the geometrical structure of the surface.
2.2. Measurement of room impulse responses

RIRs were measured in different rooms with different SMDs (100 cm, 200 cm, 300 cm). The measured reverberation times $T_{60}$ can be seen at the dots in Figure 3. The measuring of the RIRs was done with a sweep excitation signal, which increases the frequency logarithmically. As an example Figure 2 shows the spectrogram of a RIR with $T_{60} = 700$ ms.

Figure 2: Spectrogram of a RIR showing the frequency dependent reverberation. Example: Office room, $T_{60} = 700$ ms.

3. Evaluation Environment

3.1. Evaluation corpus

For this study the APOLLO corpus [6] was used for evaluation, designed to evaluate speech recognizers in a realistic kitchen scenario. The corpus consists of a set of 17 command phrases in different noise scenarios, a scenario without noise was used for this observation. A subset of 20 speakers distributed of different German dialects was chosen. Each speaker repeats the commands 3 times, which leads to a total amount of 1,020 command phrases per evaluation. The original recording sampling frequency of $f_s = 48$ kHz was down sampled to $f_s = 16$ kHz.

3.2. Speech recognizer

A command phrase recognizer developed as a part of the UASR project [7] at our laboratory was used. Its feature extraction comprises a 31 channel mel scaled filter bank, computation of velocity and acceleration features and a PCA based dimension reduction. The recognizer works with 44 monophone HMMs with automatically inferred graph topologies and Gaussian mixtures with full covariance matrices. The lexicon and language model are represented by finite state grammars. There are also rejection, noise cancellation and speaker adaptation mechanisms which, however, were all turned off for the experiments described here.

4. Experiments

For each of the following recognition experiments the related set of original (respective modified) RIRs was convoluted with the the original evaluation corpus of 1,020 command phrases. For each of the so generated subcorpora the recognition rate (RR in %) was measured. The results are illustrated in the associated diagrams.

4.1. Dependency of the RR on the reverberation time

As a starting point the influences of different rooms and different SMDs were evaluated. The strong degradation of the ASR performance with increasing reverberation times is shown in Figure 3.

Concerning the different properties of the RIR as mentioned in 2.1 the original RIRs used at six points of the graphs in Figure 3 (marked with a circle) were chosen for the following experiments with modified RIRs (refer Table 1). The modifications of the original RIRs aimed to find out the dependencies of the recognition performance on

- Early reflections
- Late reflections
- High frequency reflections
- Low frequency reflections

The modifications only apply to the reverberation phase $h_r$ of the RIRs. The direct sound impulses $h_0$ representing the clean speech signal were left unchanged.

4.2. Influence of early reflections only

To find out whether early or late reflections are most harming to speech recognition this experiment just cuts the RIR after a cut off time $t_{cut off}$ as illustrated in Figure 4. Before cutting the end was weighted with the second half of a 512 point Hamming window to smooth the cut. The cut off time measures from the end of the direct sound impulse until the end of the hamming window. Now all of the 6 chosen RIRs in Table 1 were cut at different cut off times. Figure 5 shows the result of the evaluation experiment. It can be seen that a stepwise extension between 100 ms and 300 ms is most harmful to speech recognition. Very early reflections ($< 100$ ms) and late reflections do not degrade the recognition performance so significantly.

Table 1: Original RIRs for the following sub experiments.

<table>
<thead>
<tr>
<th>Number</th>
<th>SMD</th>
<th>$T_{60}$</th>
<th>Room</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100 cm</td>
<td>400 ms</td>
<td>Living room</td>
</tr>
<tr>
<td>2</td>
<td>300 cm</td>
<td>400 ms</td>
<td>Living room</td>
</tr>
<tr>
<td>3</td>
<td>100 cm</td>
<td>700 ms</td>
<td>Office room</td>
</tr>
<tr>
<td>4</td>
<td>300 cm</td>
<td>700 ms</td>
<td>Office room</td>
</tr>
<tr>
<td>5</td>
<td>100 cm</td>
<td>3,000 ms</td>
<td>Staircase</td>
</tr>
<tr>
<td>6</td>
<td>300 cm</td>
<td>3,000 ms</td>
<td>Staircase</td>
</tr>
</tbody>
</table>
4.3. Influence of late reflections only

To study the effect of late reflections the former experiment had the disadvantage that the recognition rate was already very much degraded when late reflections were added. This experiment takes the opposite way. Adding the reflections are started from the end of the RIR as illustrated in Figure 6. Also a cut off time is introduced which describes the cut of the early reflections from the impulse response. The cut is done again with the second half of a 512 point Hamming window which reaches its peak at the maximum point of the direct sound impulse. The second cut is done with the first half of the same Hamming window which starts after $t_{\text{cut off}}$. This modification is done for different cut off times. The results in Figure 7 show again that late reflections are not as harmful as early reflections. The results of the experiment (with the exception at $T_{60} = 3,000$ ms) show that adding reflections $< 300$ ms degrade the RR much more than later reflections.

4.4. Influence of low frequency reflections only

This experiment aims to find out which frequencies of the reverberation have the strongest effect on the degradation of the RR. Therefore the reverberation phases of the six RIRs from Table 1 were low pass filtered as illustrated in Figure 9. The LP cut off frequency was increased in even mel scale steps. This procedure was chosen because the used speech recognizer analyzes the speech signal with a mel scaled filter bank. The stepwise added reverberation will therefore disturb a proportional increasing number of components of the feature vector. Thus the change of the RR from one step to another keeps comparable. The used frequencies are shown in Figure 8. The mel scale was derived following the equation

$$Z = 2,595 \cdot \log_{10} \left(1 + \frac{f}{700}\right).$$

The results in Figure 10 indicate that reverberation frequencies lower than 256 Hz do not harm the speech recognition at all. Increasing the frequency components to 1,100 Hz degrades the recognition rate down to its minimum, which is significantly less than the RR with full band reverberation. Adding frequencies above 2,680 Hz increases the recognition rate up to the full band reverberation RR (ref. Figure 3). It is assumed that the reason for this increasing effect is the energy enhancement achieved by the reverberation in this higher frequency range which leads to better recognition of fricatives in case of low frequency reverberation. Thus suppression of high frequency reverberation appears to degrade the ASR performance.
4.5. Influence of high frequency reflections only

This experiment uses the opposite way, adding high passed reverberation onto the direct impulse. Again a cut off frequency is introduced but now for the high pass filter. The HP cut off frequency was also increased in even mel scale steps following Figure 8. As in the former experiment Figure 12 shows that reverberation at frequencies above 2,680 Hz do not harm the speech recognition. Adding lower frequency reverberation degrades the recognition rate very much. This is caused by a combination of two reasons. Firstly the low frequency voiced speech has much more energy than the high frequency unvoiced speech. Secondly the reverberation is stronger for lower frequencies (ref. Figure 2).

5. Conclusions

The experiments have shown which components of the reverberation do harm speech recognition more than others. Especially low frequency reverberations approximately between 250 and 2,500 Hz do disturb the speech recognition very strongly. Higher and lower frequencies do not have a disturbing effect, even an enhancing effect was found as shown in Figure 10. Furthermore very early (< 70 ms) and very late reflections (approximately > 2/3 $T_{60}$) don’t have a great disturbing effect. Strong disturbances are caused by reflections which are shorter than 300 ms in most typical indoor living/working areas. Concludingly, a dereverberation method should preferably (or only) concentrate on reflections between 250 and 2,500 Hz and between 70 and 300 ms for common environments.

6. References