Simulation of the Hands-free Speech Input to Speech Recognition Systems by Measuring Room Impulse Responses

Hans-Günter Hirsch[1], Andreas Kitzig[1], Klaus Linhard[2]

[1]Department of Electrical Engineering and Computer Science, Niederrhein University of Applied Sciences, Krefeld
E-Mail: hans-guenter.hirsch@hs-niederrhein.de
Web: http://dnt.kr.hsnr.de/

[2] Daimler AG, Ulm, E-Mail: klaus.linhard@daimler.com

Abstract

A hardware and software approach is presented in this paper to measure the room impulse response that defines the transmission of audio signals in a room. This approach was developed within the European SpeeCon project [1].

Graphical user interfaces have been designed to estimate the room impulse response from the recordings of noise signals that are transmitted in the room and to analyse the impulse response with respect to the reverberation time and the corresponding frequency response. The impulse response can be taken to artificially create speech data that contain the influence of a hands-free speech input in a room. We used these speech data to investigate the performance degradation of a speech recognition system for this acoustic input condition. A few exemplary results are presented.

1 Measurement of room impulse response

The goal of the European SpeeCon project [1] was the collection of speech data for different languages with a focus on recording speech utterances in hands-free mode inside rooms. This should support the development of recognition systems that allow e.g. the control of electronic devices by a speech input in hands-free mode. To measure the acoustic condition in each individual recording session the hardware set-up shown in figure 1 has been developed. The intention is an estimation of the room impulse response that can be used to describe the transmission of an audio signal in a room. With the impulse response it is possible to individually analyse the acoustic condition of each recording session. Furthermore, speech data can be artificially created that contain the effect of a hands-free speech input in this specific situation.

A pink noise and a maximum length sequence (MLS) are played back from a CD player via a loudspeaker. Instead of the usually applied white noise a pink noise with an energy distribution that decreases to higher frequencies is used to compensate the frequency characteristics of the small loudspeaker. The noise signals are recorded with two microphones and stored on a PC as digital signals at a sampling rate of 16 kHz. One microphone is close to the loudspeaker. The second microphone is placed at the desired position in the room where we want to measure the impulse response. An impulse response could be estimated by comparing the recorded signal at microphone 2 with the noise signal as it is stored on the CD player. But in this case the estimated impulse response would also include the transmission characteristics of the loudspeaker. This can be avoided by the second recording close to the loudspeaker. Then, the two microphone signals can be taken to estimate the transmission characteristics between the microphones.

We apply two approaches to determine the impulse response either from the recordings of the pink noise or from the recorded MLS sequences. In case of the pink noise we can estimate the power density spectrum for each of the two microphone signals. E.g. the Welch method can be applied where the noise signal is split into segments. For each segment the spectrum is calculated with a DFT. The power density spectrum is determined as average spectrum over all segments. The ratio of the power density spectrum from microphone M2 versus the corresponding spectrum of M1 leads to an estimation of the room transfer function.

A MLS sequence has the interesting property that its autocorrelation function is approximately a Dirac impulse. This is especially true for long MLS sequences. We are using a MLS sequence of length 16383. The signal recorded by microphone M2 can be described as the convolution of the signal recorded by microphone M1 and the room impulse response.

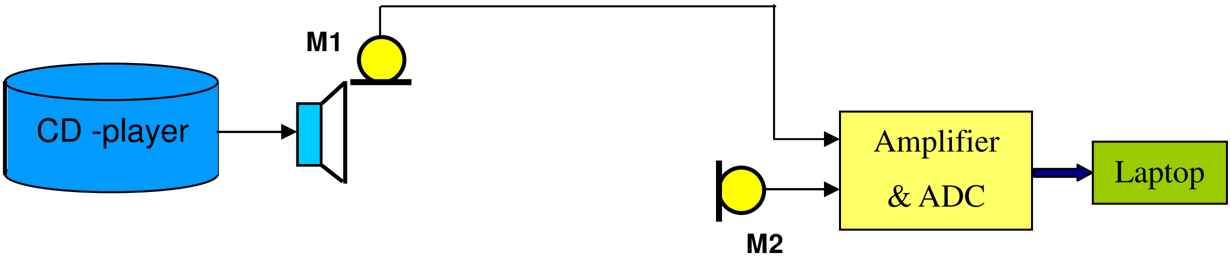


Figure 1: Hardware set-up to measure the impulse response of a room

Thus, the crosscorrelation function between the two microphone signals can be described as the convolution of the autocorrelation function and the impulse response. Due to the property of the autocorrelation function being a Dirac impulse the result of calculating the crosscorrelation between the two signals is directly the wanted room impulse response [1].

2 Graphical user interfaces to estimate and analyse the room impulse response

We designed a first graphical user interface (GUI) shown in figure 2 to estimate and visualize the room impulse response. The noise signals recorded by the two microphones and sampled at a rate of 16 kHz can be selected and loaded. The time signals are plotted in the two graphs on the left. The first noisy segment with a duration of 30 s contains the pink noise where the second segment with a duration of 12 s contains the recorded MLS sequence. After estimating the impulse responses with the two methods described in the previous section the room impulse responses are plotted in the graphs on the right together with the corresponding transfer functions. The upper graphs contain the estimation based on the usage of the pink noise where the result of comparing the MLS sequences is shown in the lower graphs. The impulse responses are fairly similar that are estimated with the different noise signals and methods. The signals shown in figure 2 have been recorded in a small conference room with a reverberation time of about 0,7 s at a distance of about 2,5 m between loudspeaker and microphone M2.

A second GUI shown in figure 3 has been designed to analyse the measured room impulse response with respect to the reverberation time and the corresponding transfer function. The room impulse response $h_{rir}(t)$ can be loaded and plotted in the upper left graph. The logarithm of the energy contour $E(\tau)=\int\limits_{\tau}^{\infty} h_{rir}^{2}(t)\,\partial t$ is calculated and plotted to estimate the reverberation time. $E(\tau)$ describes the energy contained in the impulse response from time τ till the end. Thus, $E(\tau)$ will have its maximum at the beginning of the impulse response for $\tau=0$ and will decrease towards zero for increasing values of τ. Looking at the contour of $10\cdot\log_{10}[E(\tau)]$ this should be an almost linearly decaying function.

The reverberation time is defined as the time where a decay of 60 dB is observed. This high dynamic range of 60 dB could only be measured with very good recording and conversion equipment and in an acoustic environment where almost no noise occurs in the background. Thus, this decrease down to -60 dB will not be seen in most set-ups. The example shown in figure 3 shows a linear decay of the logarithmic energy down to about -40 dB. The GUI allows the manual selection of a certain dB range. In the shown example the range between -5 dB and -40 dB is taken to fit a linearly decaying polynomial to the logarithmic energy contour in the selected dB range. The reverberation time is estimated to 0,67 s in this example based on this polynomial.

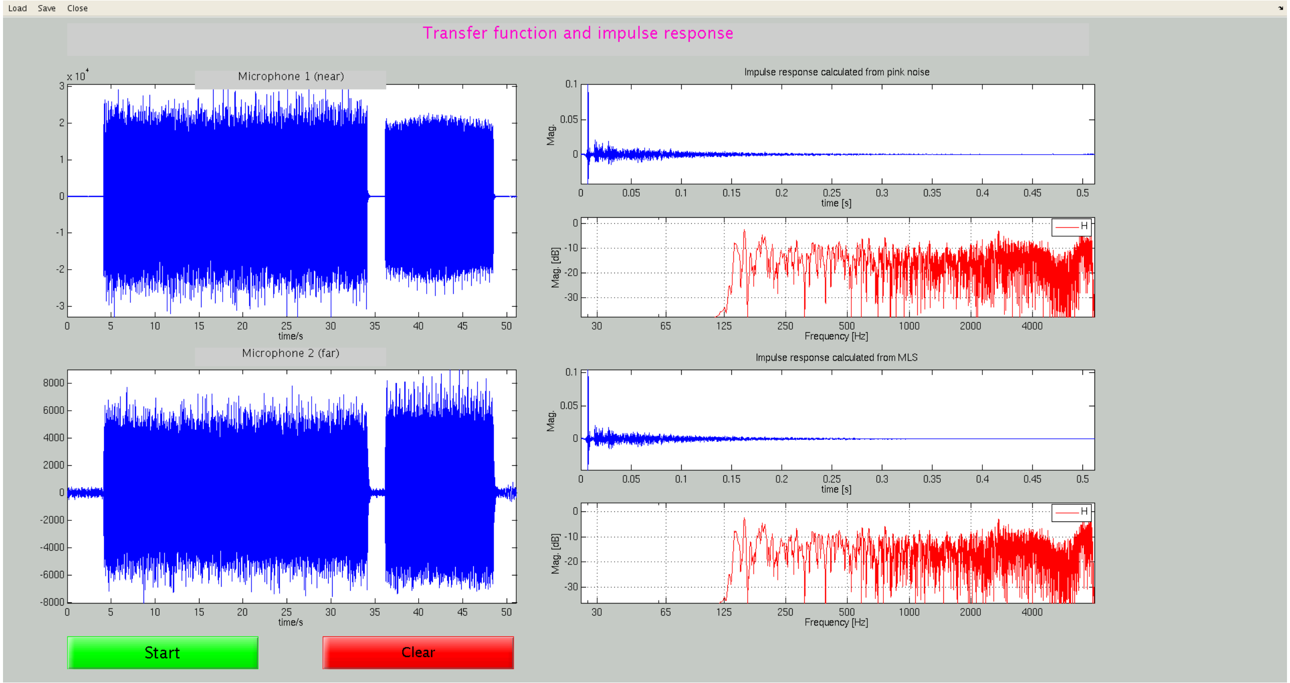
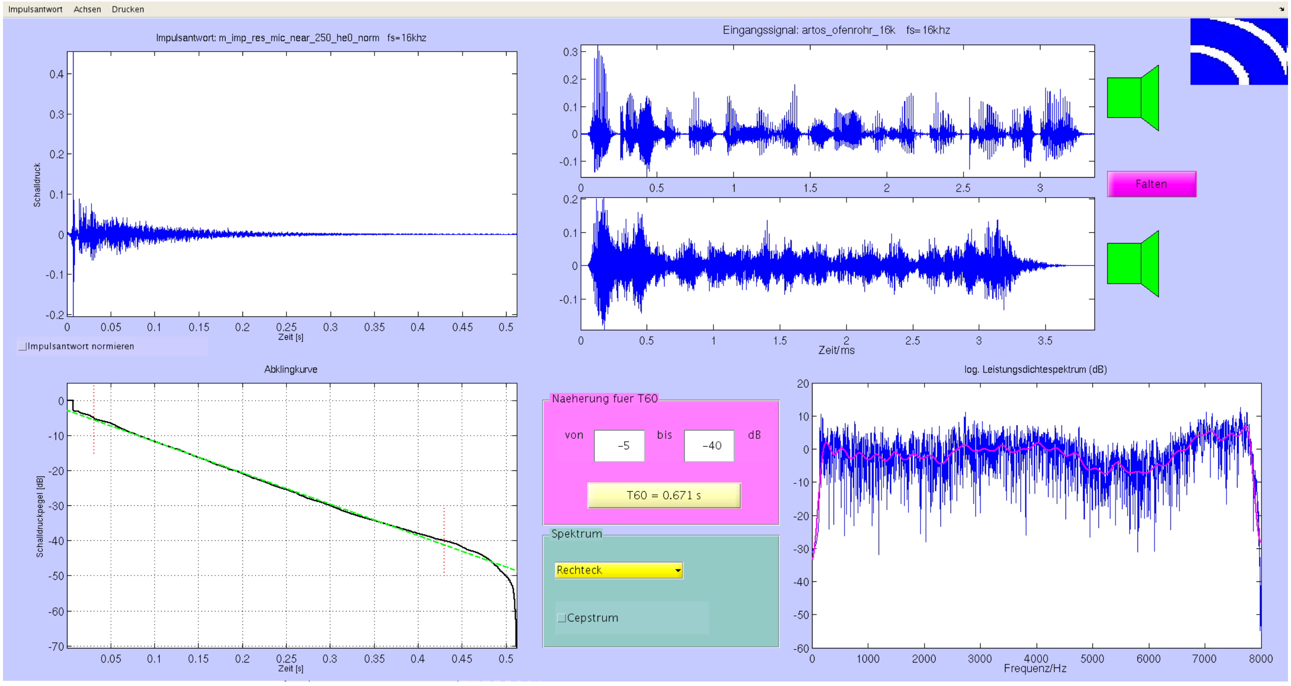


Figure 2: Graphical user interface to estimate a room impulse response from the two microphone recordings



□Impulsantwort normieren

von -5 bis -40 dB

T60 = 0.671 s

Rechteck

□Cepstrum

Figure 3: Graphical user interface to analyse a room impulse response

Furthermore the corresponding frequency response $H_{rir}(f)$ is calculated by applying a DFT. The logarithmic power density spectrum as well as a smoothed version are plotted together in a separate graph. The smoothed version is determined by applying a cepstral liftering to the logarithmic spectrum. The smoothed contour of $\log|H_{rir}(f)|^2$ is used to determine the frequency regions where the spectral components will be attenuated. This knowledge can be especially helpful when applying the rough spectral analysis of a speech recognition system. To acoustically experience the recording of an audio signal under the condition of the measurement set-up, an audio signal can be loaded. The signal is convolved with the estimated impulse response so that the user can listen to the original signal and the recording in the hands-free situation. The time signals of both versions are plotted in the upper graphs on the right together with two loudspeaker icons for listening.

3 Recognition results

The measured impulse response is taken to artificially create speech data that have been recorded in hands-free mode. A software tool [2] is available to simulate the hands-free speech input on a whole set of speech signals by convolving each signal with the impulse response. The tool can also be used to optionally add a noise signal at a desired signal-to-noise ratio (SNR) and to simulate the transmission over a mobile communication network.

We used the measurement set-up to determine two sets of impulse responses in two rooms. The first room is a small conference room with a reverberation time of about 0,7 s, the second room is an office with almost the same reverberation time. Each set contains the impulse responses for a varying distance between loudspeaker and microphone. We estimated these sets of impulse responses with the intention of investigating the influence of the distance between speaker and microphone on the performance of a speech recognition system. At a close distance to the microphone the sound on the direct path from the speaker to the microphone is dominant against the sound that reaches the microphone via reflections at the walls and e.g. the furniture in the room. The ratio between the levels of the direct and the diffuse sound in the room is taken as a quantitative measure for this effect. A so called hall radius can be roughly calculated where both levels take approximately the same value [3]. The final goal of our work is the modification of an existing approach for adapting the Hidden Markov models (HMMs) [4] to cope with the effect of the varying distance and the varying ratio between the energies of the direct sound and the stationary sound field. We made these sets of impulse responses publicly available for research purpose [5].

Figure 4 shows the word error rates for recognizing this part of the TIDigits data base [6] that contains single English digits only. We focused on the recognition of isolated digits first because the influence of the reverberation on fluently spoken sequences of digits is more complex. The error rates are shown dependent on the distance between speaker and microphone in the conference room. The hall radius takes a value of approximately 45 cm in this room. A "standard" cepstral analysis has been applied to extract 12 cepstral coefficients and the logarithmic energy together with the corresponding Delta and Delta-Delta coefficients as acoustic features. Gender dependent HMMs with 16 states and a mixture of 2 Gaussians per state are used to model the 11 English digits including "zero" and "oh". The word error rate decreases only a bit for distances below the hall radius where a larger deterioration is observed for distances above the hall radius.

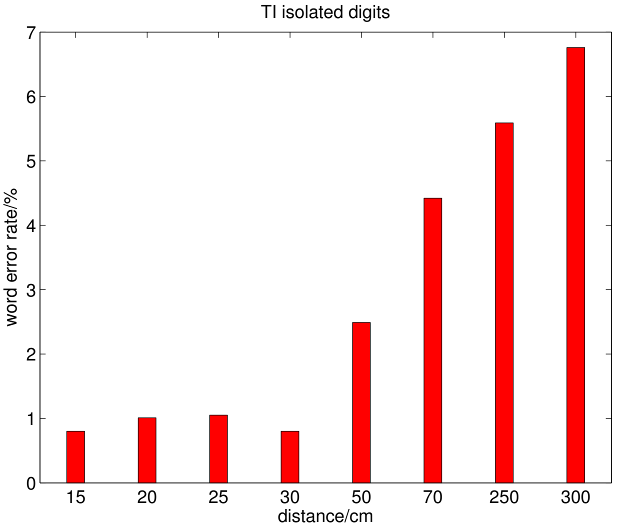


Figure 4: Word error rates for the TI isolated digits dependent on the distance speaker – microphone

For another set of experiments we selected about 4800 utterances from the German speech corpus “RVG” (regional variants of German) [7] containing sequences of German digits with a total of about 20000 digits. We intend to investigate the influence of a hands-free speech input on the performance of different recognition systems in an acoustic environment with noise in the background.

	SNR/dB		
	15	10	5
noise only	4,61 %	7,22 %	14,13 %
noise & hands-free	6,02 %	8,96 %	16,41 %

Table 1: Word error rates for noisy versions of the RVG subset in a car environment

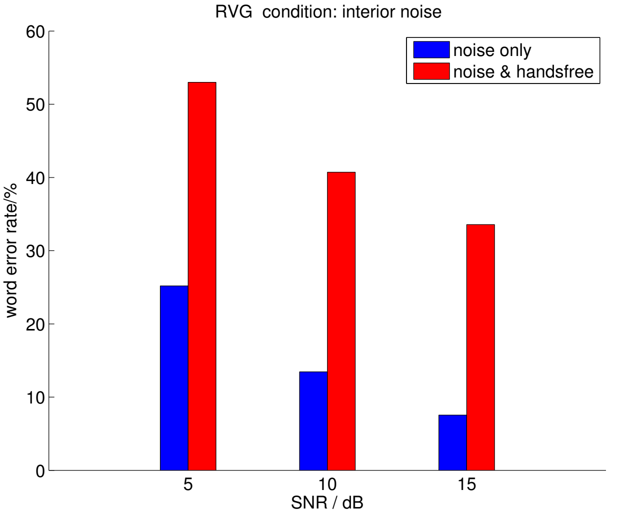


Figure 5: Word error rates for noisy versions of the RVG subset in a room environment (T$_{60}$ ~ 0,7s)

A room impulse response measured in a car (VW Touran) and a noise signal recorded inside a car have been used to create several versions of the RVG subset with and without considering the input in hands-free mode. Applying a robust feature extraction scheme [8] and using whole word HMMs that have been trained on the original data, the word error rates listed in table 1are achieved. It turns out that the additional performance degradation due to the hands-free input is fairly low. The reverberation time of the car body takes a small value less than 100 ms which is the reason for the low increase of error rates. Taking the impulse response measured in the conference room with a reverberation time of about 0,7 s at a large distance between speaker and microphone and adding noise signals as they typically occur inside rooms the word error rates are obtained that are presented in figure 5. The speech transmission in a room with a much higher reverberation time as in the car causes a considerable increase of the error rate.

The applied feature extraction scheme [8] contains a processing block for noise reduction based on a two stage Wiener filtering. This type of adaptive filtering needs an estimation of the noise spectrum assuming a stationary background noise. The influence of a hands-free speech input can not be compensated with this filtering approach because the reverberation in a room has mainly the effect of modifying the temporal structure of speech. Due to sound reflections at the walls and the furniture the acoustic features of speech segments will be seen for a longer period of time. This leads to the superposition of the acoustic features from consecutive segments. The existing schemes for extracting robust acoustic features do not compensate such effects. They have to be extended by an additional processing to keep the performance high in case of a hands-free speech input. But so far, no approaches for de-reverberation are known that could be easily integrated in the existing feature extraction.

References

[1] K. Linhard. Measurement of room acoustics and noise characteristics, SpeeCon project report, available at http://www.speechdat.org, 2002

[2] H.-G. Hirsch, H. Finster. The simulation of realistic acoustic input scenarios for speech recognition systems, Eurospeech Conference, Lissabon, Portugal, 2005

[3] H. Kuttruff. Room Acoustics, Spon Press, 2000

[4] H.G. Hirsch. Automatic speech recognition in adverse acoustic conditions, in *Advances in Digital Speech Transmission*, John Wiley and sons, 2008

[5] http://dnt.kr.hsnr.de/ (➔ download section)

[6] R.G. Leonard, A Database for speaker-independent digit recognition. *Proc. of ICASSP*, Vol. 3, p. 42.11., 1984

[7] S. Burger, F. Schiel. RVG1 – A database for regional variants of contemporary German, available at http://www.phonetik.uni-muenchen.de/Bas/

[8] ETSI standard document. Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Advanced Front-end feature extraction algorithm; Compression algorithm. ETSI document ES 202 050 v1.1.3 (2003-11), Nov. 2003.