The Simulation of Realistic Acoustic Input Scenarios for Speech Recognition Systems

H. Günter Hirsch, Harald Finster

Niederrhein University of Applied Sciences, Krefeld, Germany
{hans-guenter.hirsch},{harald.finster}@hs-niederrhein.de

Abstract

A tool for simulating the acoustic conditions during the speech input to a recognition system and the transmission in telephone networks is presented in this paper. The simulation covers the hands-free speech input in rooms and the existence of noise in the background. Furthermore the presence of telephone frequency characteristics can be simulated. Finally the transmission in a cellular telephone system like GSM or UMTS is covered including the encoding and decoding of speech and the transmission over the erroneous radio channel.

The tool has been realized by integrating functions from the ITU software library for implementing telephone frequency characteristics and the estimation of the speech level as well as software modules from ETSI and 3GPP for the AMR encoding and decoding of speech.

A Web interface has been designed to experience the simulation tool with acoustic examples.

1. Introduction

The missing robustness to different acoustic environments is still one of the key issues that prevent the usage of speech recognition systems in a lot of real world scenarios. Most of the research work in this area has focused on the influences of additive background noise and of unknown stationary frequency characteristics. But there are further effects that have influence on the speech signal. The usage of a speech recognition system is especially advantageous in situations where a person can not make use of it's hands. A typical example is a driver in a car who wants to control devices in the car or wants to retrieve information from a remote system by voice. The acoustic environment has a major influence on the speech and on the recognition performance in such situations of a hands-free speech input. In case the user is not wearing a close talking microphone there will be not only distortions due to noise in the background. The speech input takes place in a reverberant environment where a lot of reflections add up to the original speech signal. This effect has been only considered in a few investigations so far.

A further typical scenario for using a recognition based speech dialogue system is the communication with a mobile phone while not having access to an Internet terminal for information retrieval. The speech signal is encoded and decoded for transmission in mobile networks. Furthermore the erroneous cellular channel has an influence on the speech signal and on the recognition performance.

The goal of this work is the creation of a tool to simulate all of the above mentioned effects. This tool will be described in the following section including the applied signal processing techniques. Furthermore a Web interface is presented that has been set up to let people also acoustically experience this simulation tool.

2. Simulation tool

Figure 1 gives an overview about all effects of the acoustic environment in practical situations where applying a speech recognition system.

2.1. Hands-free speech input in rooms

In case of not using a close talking microphone the acoustic environment of the room modifies the speech signal

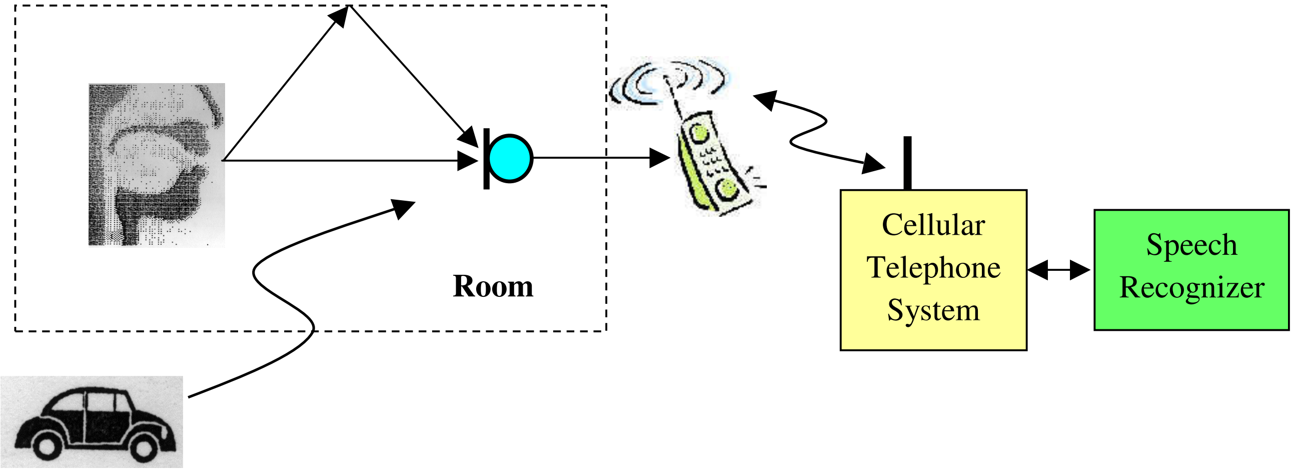


Figure 1: Speech input and transmission scenario for recognition based speech dialogue systems

by adding multiple delayed, attenuated and filtered versions of the signal itself caused by multiple reflections at the walls. This effect is called reverberation. It can be described by the impulse response for the transmission between the positions of the speaker and the microphone. In practice this is usually a time variant function when the speaker moves in the room or the room configuration is modified by e.g. opening or closing doors or windows.

The focus of this work is on situations as they are of interest for practical applications of speech recognition systems. Three situations are considered as being of high interest:

- hands-free speech input while driving a car
- hands-free speech input at a desk in an office room with the intention of controlling the phone itself or using it for information retrieval from a remote system
- hands-free speech input in a living room with the intention of controlling e.g. audio or video devices

We created three impulse responses that describe the stationary situation with speaker and microphone at fixed positions.

First reflections for the office room and the living room are simulated with a mirror image model up to 9th order [1]. Different approaches have been investigated to add further late reflections [2],[3]. We ended up by applying an own approach where we spread the impulse response of the mirror image model over a longer time period. We extracted the late reflections from the spread impulse response and combined it with the early reflections of the mirror image model. This leads to a quite natural sounding reverberation. The coefficients of the impulse response have been weighted to represent a room with a reverberation time of approximately 0,4 s for the office room and of approximately 0,6 s for the living room. Furthermore the possibility has been introduced to vary the reverberation time in a certain range.

Because the simulation of a hands-free communication in a car is not that easy we took a measured impulse response. For the measurement a loudspeaker was placed at the position of a driver’s mouth and the microphone in the neighborhood of the interior mirror inside a car.

The three impulse responses are shown in figure 2 together with their corresponding frequency responses.

2.2. Additive background noise

The presence of noise in the background can be simulated by adding a recorded noise signal to the speech signal at a desired SNR (signal-to-noise ratio). The estimation of speech and noise levels is done according to ITU recommendation P.56 [4]. The same approach has been applied in earlier

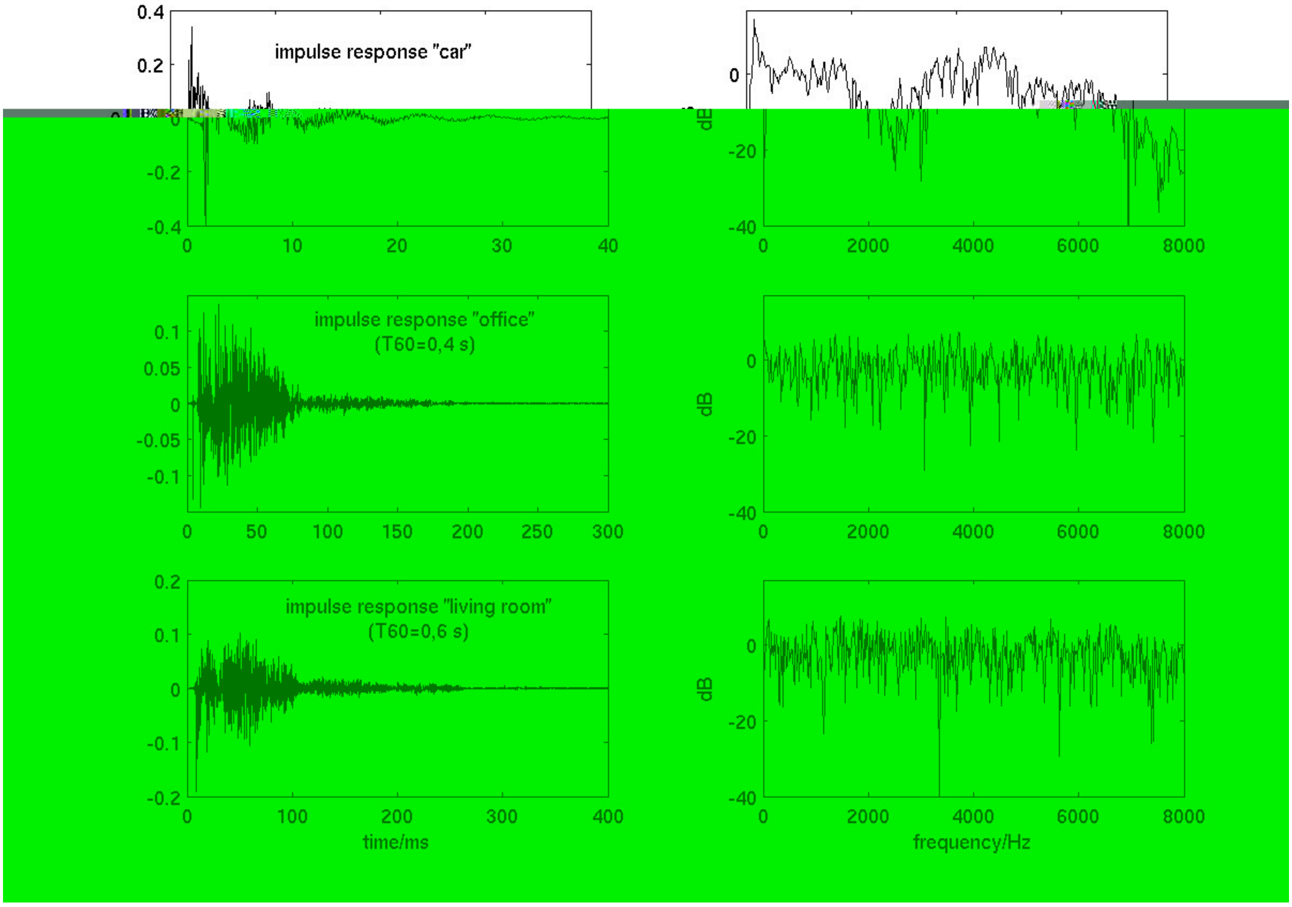


Figure 2: Impulse and frequency responses of three hands-free speech input scenarios

investigations where noisy speech data has been generated for the evaluations inside the ETSI working group Aurora [5]. The speech and the noise signals are filtered with a G.712 frequency characteristic beforehand for estimating their levels. Thus the SNR is related to the energy of the signals in the frequency range from about 300 to 3400 Hz. Especially the misleading determination of a too bad SNR is considerably improved with this filtering as it occurs for noise signals with frequency components of high energy below 300 Hz.

2.3. Telephone frequency characteristics

Some frequency characteristics can be applied to simulate the recording with a telephone device and the transmission over telephone networks. The applied frequency responses have been defined by ITU [4] and are well known by their abbreviations G.712, IRS and MIRS. In all cases frequency components below 300 Hz and above 3400 Hz are considerably attenuated. The G.712 filtering has a flat characteristic in the range between 300 and 3400 Hz where the frequency responses of IRS and MIRS show an increasing trend in this range with a slightly higher attenuation at low frequencies.

2.4. Transmission over cellular telephone networks

The usage of mobile phones is a typical scenario while accessing a speech dialogue system for information retrieval. Encoding and decoding of the speech as well as the transmission over the erroneous cellular channel modify the speech signal. The influence of these distortions can be avoided by the approach of a distributed speech recognition where the acoustic features are extracted in the terminal and transmitted as data in a quite save mode. But so far this approach is not seen in practical applications. Thus it seems to be useful to simulate also the transmission over cellular networks. The AMR (adaptive multi rate) coding schemes are applied for considering the influence of speech encoding and decoding. These schemes are and will be mainly used for speech transmission in GSM and UMTS networks. There exist two sets of coding schemes for encoding speech in the narrow-band frequency range up to 4 kHz and in the wide-band range up to about 7 kHz. The AMR-NB (narrow-band) codec includes 8 coding modes with data rates between 4,75 and 12,2 kBit/s. The AMR-WB (wide-band) coding scheme includes 9 coding modes with data rates between 6,6 and 23,85 kBit/s.

The influence of the transmission over GSM and UMTS channels is simulated by applying bit error patterns to the data stream between speech encoding and decoding. These error patterns have been derived by simulating channel encoding and decoding together with the “standardized” error patterns that are applied between channel encoding and decoding. The error patterns applied on the data stream after channel encoding have been created for typical transmission scenarios as e.g. driving in a car. As advantage of creating “shifted” versions of the typical bit error patterns that can be directly applied to the data stream after speech encoding, no channel encoding and decoding blocks are needed in the simulation tool. In case of GSM transmission, error patterns exist for all AMR coding modes that are designated for their usage in GSM networks. For each speech coding mode there exist patterns for different C/I (carrier-to-interference) ratios. The value of the C/I in dB describes the quality of the cellular channel. E.g. a value of 4 dB describes the communication at the border of a radio cell where a value of 16 dB describes the situation in the center of the radio cell.

In case of the CDMA based transmission in UMTS networks the quality of the cellular channel is defined by the frame error rate that describes the percentage of erroneous frames. Error patterns are available for all AMR coding modes and frame error rates of 0,5%, 1% and 3%.

2.5. Realization

The simulation tool has been realized as C++ program including a lot of functions written in C as they are available in the ITU software library respectively as exemplary code for the speech coding schemes from ETSI and 3GPP [6],[7]. It has been tested by running a lot of recognition experiments on data that have been artificially distorted with this tool. The recognition results are presented in a separate paper on this conference.

3. WWW presentation

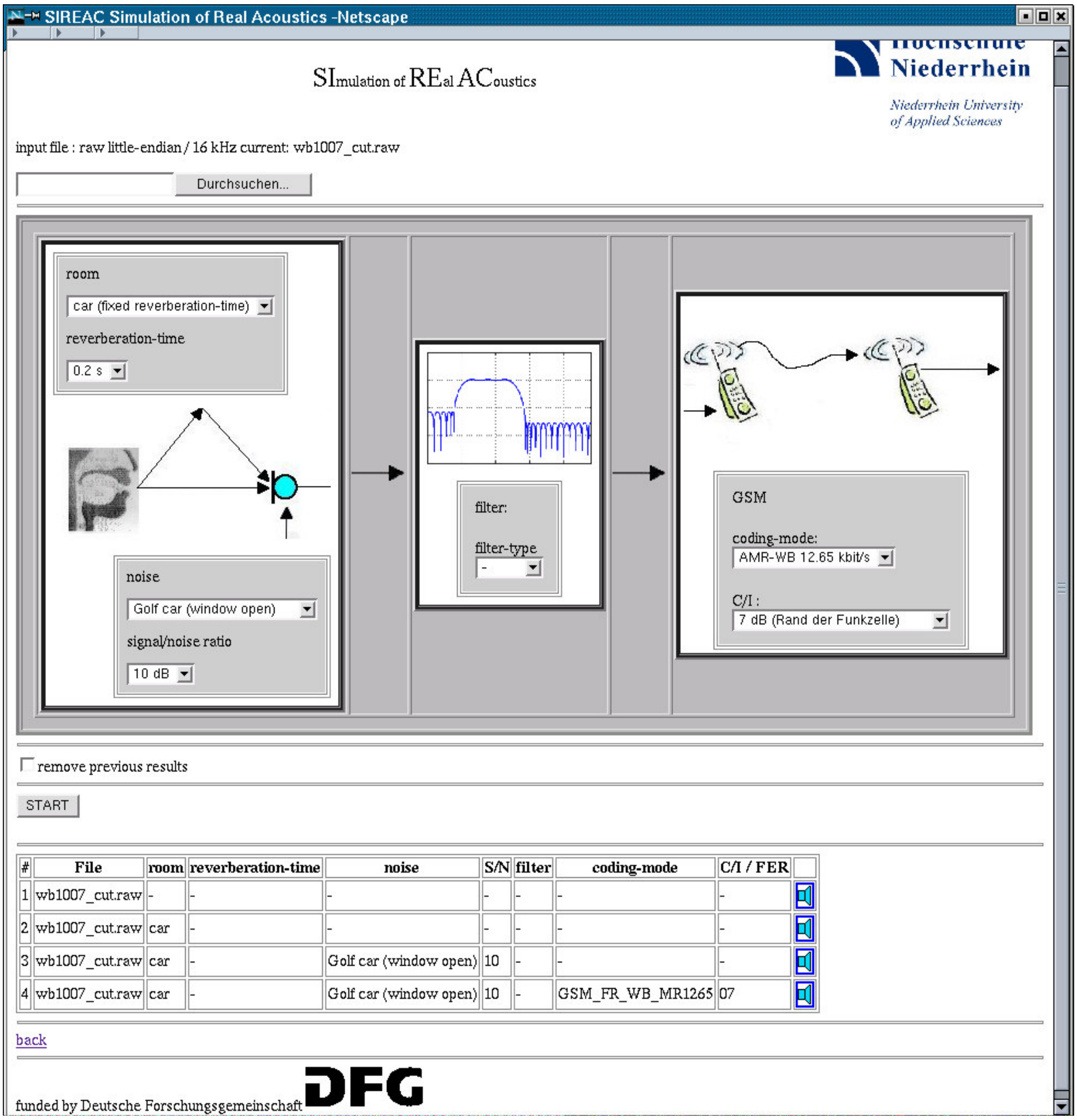
To publicly present this work and open the possibility of acoustically experiencing the conditions of a speech input in practical situations we created an interface in the World Wide Web that can be accessed over the Internet [8]. Figure 3 shows the graphical interface of the main page.

It is possible to define a desired scenario by selecting certain conditions in the pull down menus. The user can select an own speech file available on his client system. If the user has no speech file available on his client system, a short speech utterance available on the host system will be taken for the simulation instead. When the simulation is started, the selected speech file will be transmitted to the host system. After processing a further line will appear in the table of processed conditions. Pressing the mouse button on the loudspeaker symbol enables the immediate listening to the processed speech signal or storing the speech signal on your client system. The table of processed conditions allows the easy acoustic comparison of different scenarios.

4. FaNT – Filtering and Noise adding Tool

As mentioned above, this tool can also be seen as the extension of an earlier tool called FaNT, which was used to create the artificially distorted speech data for the widely used Aurora-2 corpus [5]. FaNT can be used to add noise at a desired SNR and to apply certain telephone frequency characteristics. It has now been made available for public download at http://dnt.kr.hsnr.de/download.html .

In order to define SNR in a way relevant to human speech perception, FaNT allows frequency weighting according to A-weighting or telephone frequency characteristics for estimating the levels of the speech and noise signals. To avoid being misled by speech pauses, the levels of the signals are calculated according to ITU recommendation P.56. FaNT can help make the results of different researchers more comparable by providing a shared definition for SNR. It also allows the same noisy version of an original speech corpus to be recreated at different sites without distributing the noisy data itself, which can be helpful for copyright reasons. Some comments about the problem of defining SNR are available at the download page.



SIREAC Simulation of Real Acoustics -Netscape

SImulation of REal ACoustics

input file : raw little-endian / 16 kHz current: wb1007_cut.raw

room: car (fixed reverberation-time)

reverberation-time: 0.2 s

noise: Golf car (window open)

signal/noise ratio: 10 dB

filter:

filter-type: -

GSM

coding-mode: AMR-WB 12.65 kbit/s

C/I : 7 dB (Rand der Funkzelle)

□ remove previous results

#	File	room	reverberation-time	noise	S/N	filter	coding-mode	C/I / FER	
1	wb1007_cut.raw	-	-	-	-	-	-	-	🔈
2	wb1007_cut.raw	car	-	-	-	-	-	-	🔈
3	wb1007_cut.raw	car	-	Golf car (window open)	10	-	-	-	🔈
4	wb1007_cut.raw	car	-	Golf car (window open)	10	-	GSM_FR_WB_MR1265	07	🔈

funded by Deutsche Forschungsgemeinschaft DFG

Figure 3: Web interface for experiencing the simulation tool (http://dnt.kr.hsnr.de/sireac.html)

5. Acknowledgements

The authors would like to thank Andreas Silzle from the Institute of Communication Acoustics at the Ruhr-University in Bochum for generating the impulse responses of the two rooms with their IKA-SIM system [1]. Furthermore we would like to thank Gerald Enzner from the Institute of Communication Systems and Data Processing at the University of Technology in Aachen for providing us with themeasurement of the impulse response in the car.

Special thanks address the Deutsche Forschungsgemeinschaft for funding this project.

6. References

[1] Silzle, A., Novo, P., Strauss, H. "IKA-SIM: A System to Generate Auditory Virtual Environments", *116th Convention of the Audio Engineering Society*, Berlin, Germany, 2004.

[2] Moorer, J. A. “About this Reverberation Business”, *Computer Music Journal*, 1979, pp. 13-28.

[3] Jot, J.-M. “An Analysis/Synthesis Approach to Real-time Artificial Reverberation”, *Proceedings of ICASSP*, San Francisco, USA, 1992.

[4] http://www.itu.org

[5] Hirsch, H.G., Pearce, D. “The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions”, *ISCA workshop ASR2000*, Paris, France, 2000

[6] http://www.etsi.org

[7] http://www.3gpp.org

[8] http://dnt.kr.hsnr.de/sireac.html