A MATLAB Simulation of “Shoebox” Room Acoustics for use in Research and Teaching

Douglas R. Campbell[1]*, Kalle J. Palomäki*[2] *and Guy J. Brown*[3]

[1]School of Information and Communication Technologies, University of Paisley,
High Street, Paisley PA1 2BE, Scotland, U.K.
[2]Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing,
P.O. Box 3000, FIN-02015 HUT, Finland.
[3]Department of Computer Science, University of Sheffield,
211 Portobello Street, Sheffield S1 4DP, U.K.

d.r.campbell@paisley.ac.uk kalle.palomaki@hut.fi g.brown@dcs.shef.ac.uk

Abstract

A simulation of the acoustics of a simple rectangular prism room has been constructed using the MATLAB m-code programming language. The aim of this program (Roomsim) is to provide a signal generation tool for the speech and hearing research community, and an educational tool for illustrating the image method of simulating room acoustics and some acoustical effects. The program is menu driven for ease of use, and will be made freely available under a GNU General Public Licence by publishing it on the MATLAB Central user contributed programs website. This paper describes aspects of the program and presents new research data resulting from its use in a project evaluating a binaural processor for missing data speech recognition.

1. Introduction

The Roomsim program is a simulation of the acoustics of a simple rectangular prism room constructed using the MATLAB m-code programming language. The foundation on which Roomsim is built is the publication of a Fortran routine by Allen and Berkley [1] in 1979 that has been translated into various computer languages to use as the core of simulations of “shoebox” room acoustics. The reason for producing this program is our belief that a freely available and enhanced MATLAB implementation will be of advantage to digital signal processing (DSP) researchers and educators working in areas of acoustic signal processing, speech and hearing, because of their familiarity with MATLAB, which has arguably become the prime DSP algorithm development platform.

The image method of simulating room acoustics is often used as a means of generating signals incorporating “sufficiently realistic” reverberation and directional cues for the testing of audio/speech processing algorithms [2]. Commercially available programs for simulating architectural acoustics can be used for these purposes, but the cost and degree of geometric and acoustic realism is usually overkill.

Many signal processing researchers now use the MATLAB technical computing language to develop their algorithms because of its ease of use, powerful library functions and convenient visualisation tools. The authors had separately searched for a Windows PC and Unix MATLAB implementation of room acoustics that would satisfy their experimental requirements for quickly and easily generating binaural speech signals sufficiently realistic for the assessment of signal processing algorithms aimed at improving the performance of speech recognisers, hands-free telephones and hearing aids. A couple of candidate programs were found, but neither incorporated the majority of features required. Palomäki *et al*. [3] developed a MATLAB program to satisfy the immediate needs of a part of the SPHEAR project at Sheffield investigating a binaural processor for missing data speech recognition. Following the presentation of results from that project in 2002 at a Binaural Hearing workshop, Campbell was supplied with their code and used that as the core of the Roomsim program reported here.

The authors intend releasing the program as free software, under the GNU General Public Licence, to encourage its dissemination and development. The m-code source files, execute within MATLAB on a PC or Unix platform, and a compiled stand-alone executable for non-MATLAB PC installations, will be posted on a few appropriate web sites including the file exchange at MATLAB Central http://www.mathworks.com/matlabcentral/fileexchange. The existence and location of the code will be more widely disseminated via notices to various special interest groups in signal processing, binaural hearing, speech processing, audio processing and acoustics.

This paper consists of two main parts: a description of some features and facilities of the Roomsim program including areas for educational use, and a report of new experimental data from a project investigating the robust ASR system of Palomäki *et al*. [3] and extending the evaluation of it to include a number of reverberant conditions, simulated using the Roomsim program.

2. The Roomsim program

The program simulates the geometrical acoustics of a perfect rectangular parallelepiped room volume using the image-source model to produce an impulse response from each omni-directional primary source to a directional receiver system that may be a single sensor, a sensor pair or a simulated head. The method of generating an impulse response from a primary source to a receiver sensor by the method of image sources is very adequately described in reference [1] and will not be repeated here.

The Allen and Berkley algorithm has been extended to include frequency dependent sound absorption coefficients. Using this, the Roomsim program provides the ability to select different materials for each of the six major surfaces of the simulated room from a list of (at present) 20 standard building materials (plus 4 synthetics including anechoic) that have had their frequency dependent absorption coefficients tabulated.

Another desired feature was the ability to incorporate a range of receiver systems such as single sensor (e.g. mono microphone), sensor pair (e.g. two element microphone array) and simulated human head. The single and dual sensor receivers can be configured as directionally sensitive and the interpolation process recommended by Peterson [4] is incorporated. The simulation of the head utilises the Head Related Transfer Function (HRTF) data, actually Head Related Impulse Response (HRIR) data, provided from measurements made on a KEMAR mannequin at the MIT Media Lab. [5], and on 42 human subjects and a KEMAR at the Center for Image Processing and Integrated Computing (CIPIC), University of California, Davis [6]. Both sources have agreed to the distribution of their data within the Roomsim package.

A core aim was to provide user interaction having a suitable balance between flexibility for researchers and ease of use for undergraduate students. The m-code has been written to: (1) allow operation on PC and Unix platforms; (2) avoid the need for specialist MATLAB toolboxes; (3) avoid functions that cannot successfully be compiled to produce stand-alone executable code. Thus, only a basic MATLAB installation is required to run and develop the m-code, and users without a MATLAB installation may run the executable version on a Windows PC.

The program was developed on a medium specification PC (1.5 GHz Intel Pentium4, 512 MB RDRAM, 30GB HDD UDMA 100 IDE, AGP Graphics) running Windows 2000 with MATLAB v 6.5 rev. 13 installed. On this system (using an example from [1]) simulating a single source and single sensor in a room of size 3*4.6*3.8 m and RT_{60}=0.32 s, the core calculation of 28,500 images took approximately 42.5 s i.e. ≈1.5 ms per image. It has been run on desktop and notebook PCs down to 200 MHz clock frequency with 128 MB RAM. Reflection orders >10, impulse responses > 10,000 samples and multiple sources will all contribute to increasing the computational load resulting in a longer run time, especially of the core impulse response calculation and the 2D and 3D graphics.

2.1. Roomsim operation

In operation, the user specifies the dimensions of the room, its surface materials and the location of the receiver system and the primary source(s). This can be done interactively through the menu prompt system, by submitting a Microsoft Excel spreadsheet form, or by selecting a MATLAB *.mat file which saved a configuration from a previously run. The program performs various “sanity” checks and an estimate of the reverberation time (RT_{60}) of the simulated room is used to size the impulse response to be computed.

The image-source to receiver responses are then computed using the method of images, modified to take account of the variation of surface absorption coefficient with frequency and for the path length (1/R effect). The frequency dependent attenuation due to air is included if desired. If a simulated head has been selected the response from each quantised image-source direction is convolved with the relevant HRIR data. The individual image-source responses are then accumulated to form the complete pressure impulse response from each primary source to the receiver and the results plotted and saved to file.

These two-channel impulse response file(s) can then be convolved within the program with the users’ own monophonic audio files (*.wav or *.mat format). The resulting monaural, stereo, or “binaural response” can be saved as an audio or MATLAB file. If saved in *.wav format it can then be played using a Windows PCM WAV format compatible media player or sound editor. The user may sum the response files related to each simulated primary source produce the combined acoustic signal at each sensor/ear e.g. to create speech and noise signals at different locations with their appropriate reverberation, or a “cocktail party”.

2.2. List of parameters and facilities

Roomsim: User configurable parameters
Humidity of air (modifies air absorption coefficient).
Temperature of air (modifies the speed of sound).
Sampling frequency (at present 8 kHz to 44.1 kHz).
Enclosure dimensions (Lx, Ly, Lz).
Surface absorption for each of the room’s six surfaces (at present 24 frequency dependent surface types).
Air absorption model may be ON or OFF.
Receiver coordinates (x, y, z).
Receiver type (single sensor, sensor pair, HRTF).
Sensor directionality (azimuth and elevation).
Sensor separation for sensor pair.
HRTF from MIT KEMAR or CIPIC subjects.
Multiple sources, location specified as polar wrt. receiver.
Order of reflections.
Length of impulse response.
Interpolation filter (used in sensor pair case).
High-pass filter cut-off, for reduction of DC bias and low frequency ripple.
Effect of distance (1/R) may be ON or OFF.

Roomsim: Display options
Plot of surface absorption vs. frequency.
3D display of room, receiver and sources geometry.
Plot of mean reverberation time (RT_{60}) vs. frequency.
Plot of impulse response(s) vs. time or sample number, Colour coded for left (L) and right sensor (R).
Magnitude spectrum corresponding to above impulse response(s), FFT length selectable, L/R colour coded and HRTF superimposed when MIT or CIPIC data selected.
2D zoom and rotatable plan view of room, receiver, source(s), surrounding image rooms, and image sources, with source intensity indicated by L/R colour code.
3D version of the above plan view with source intensity displayed as stem plot height, L/R colour coded.
3D zoom and rotatable view of room, receiver, source(s), surrounding image rooms, and image sources, with source intensity indicated by L/R colour code.

Roomsim: File formats
MATLAB *.mat
Microsoft Windows PCM *.wav

Roomsim: Utilities for processing data files
A convolution operation that accepts both *.mat and *.wav format, and displays the impulse response data, audio file data and convolution result as line graphs, and additionally the latter two as spectrograms.
An accumulation function for use in building “cocktail party” scenarios, also with signal displays.
A converter between *.wav and *.mat, one and two channel file formats.
A high-pass filter (4th order Butterworth).
A low-pass filter (4th order Butterworth).

2.3. Educational use

The Roomsim program may be of interest to educators delivering courses or modules in acoustic signal processing, music technology, auditory perception, psychoacoustics, environmental acoustics, and digital signal processing. It can form the basis for standard demonstrations, exploratory classroom laboratories or undergraduate projects. For example, since the user has control of surface absorption, the image source method can be effectively illustrated by adding room surfaces one at a time and controlling the order of reflections displayed. The effect of surface and air absorption can be demonstrated and their dependence on room area, volume and air humidity. The simulation may be used as a reverberation chamber for generating audio effects or demonstrating the acoustic complexity of reverberant enclosures. By placing a sensor close to a reflective surface the acoustic comb filter effect can be demonstrated and the expected notch frequencies confirmed in the spectral magnitude plot. The difference in sound quality and localisation between a stereo microphone pair and a binaural “head” can be demonstrated. The 3D displays (which can be rotated and zoomed) of room geometry, receiver and primary source locations, image rooms and image sources (location and strength) are powerful visualisations of the complexity of the reverberant environment.

3. Research use

The simulation of room acoustics is an essential part of the evaluation of robust algorithms for automatic speech recognition (ASR). Palomäki *et al*. [3] have recently described an approach to robust ASR that is motivated by two aspects of human hearing; binaural processing and masking. Binaural processing is known to play an important role in the perceptual segregation of a target sound from interfering sounds, and confers resistance to the adverse effects of reverberation on speech perception. Speech perception is known to be robust even when parts of the speech signal are masked by other sounds. This leads to the notion of a ‘missing feature’ approach to ASR [7] in which a time-frequency mask is derived which partitions the acoustic spectrum into reliable and unreliable regions which are then treated differently in the ASR (see section 3.4).

3.1. Automatic speech recognition system

The ASR system consists of an auditory model front-end, divided into monaural and binaural pathways, and a “missing feature” speech recogniser. The monaural pathway is responsible for peripheral auditory processing, and for producing feature vectors for the speech recogniser. The binaural pathway is responsible for sound localisation and separation according to common azimuth.

3.2. Monaural pathway

In the first stage of the monaural pathway, the direction dependent filtering effects of the pinna, head, torso and the room interaction are modelled by convolving the acoustic input with a head-related room impulse response (HRRIR) for each ear. HRRIRs are generated using Roomsim as described in Section 2.1. Cochlear processing is simulated by an auditory filterbank consisting of 32 bandpass gammatone filters with centre frequencies spaced on the equivalent rectangular bandwidth (ERB) scale between 50 Hz and 8 kHz. To represent auditory nerve activity the outputs of gammatone filters are half-wave rectified and compressed. The auditory nerve response is passed to a model of binaural processing, which is described in the next section.

In order to produce features for the recogniser the instantaneous Hilbert envelope is computed at the output of each auditory filter [8]. The envelope is then smoothed by a first-order low-pass filter (8 ms time constant), cube root compressed and sampled at 10 ms intervals. This yields ‘rate maps’ for the left and the right ear, which are averaged. The spectral distortion caused by HRRIR filtering is compensated by a spectral energy normalisation scheme [9].

3.3. Binaural pathway

The binaural model estimates the azimuth of sound sources from their interaural time differences (ITDs), which are computed from a cross-correlogram of the auditory nerve response. In order to emphasise direct sound over the reverberant sound field we use a model of the precedence effect. This emphasizes acoustic onsets and inhibits reflections prior to the cross correlation analysis. Computing the instantaneous envelope from each gammatone filter produces an inhibitory signal, which is delayed and smoothed by a low-pass filter (15 ms time constant). The inhibitory signal is then subtracted from the simulated auditory nerve response. Because the inhibitory signal is delayed, abrupt onsets corresponding to the direct sound are unaffected, whereas later reflections are attenuated [3].

ITD-based azimuth estimates are produced as follows. Let the precedence processed left and right ear auditory nerve responses be $l(i,j)$ and $r(i,j)$, where i is the channel index and j is the time step. Then the cross correlation for delay τ is:

$$C(i, j, \tau) = \sum_{k=0}^{M-1} l(i, j-k) r(i, j-k-\tau) w(k) \qquad (1)$$

Here w is a rectangular window of M time steps. In the simulations $M=600$ is used, which corresponds to a window duration of 30 ms. Values of τ lie between ±1 ms.

Computing $C(i,j,\tau)$ for each channel i gives a cross-correlogram, which is computed at 10 ms intervals. Each cross-correlation function is mapped to an azimuth scale using a lookup table, giving a modified function $C(i,j,\phi)$, where ϕ is the azimuth in degrees.

The correlogram $C(i,j,\phi)$ is further processed by forming a “skeleton” for each channel, in which each peak in the cross-correlation function is replaced with a narrow Gaussian with a width inversely related to the channel centre frequency [3]. Thus the peaks become narrower with increasing frequency, a process similar in principle to lateral inhibition, which leads to sharpening of the cross-correlogram. Finally, the skeleton cross-correlation functions are summed over frequency to give a pooled cross-correlogram, in which the location of each sound source is indicated by a clear peak.

3.4. Missing feature speech recogniser

The speech recogniser used here employs the missing feature technique [7], in which a HMM-based speech recogniser is adapted to deal with missing or unreliable features. Generally, the classification problem in probabilistic speech recognition can be stated as the assignment of an acoustic observation vector v to a class C. Due to the application of Bayes’ rule this classification can be performed by maximising the likelihood $f(v|C)$. However, in the presence of noise or reverberation, some components of v will be unreliable or missing and the likelihood cannot be computed normally. Using the ‘missing feature’ paradigm this is addressed by partitioning v into reliable and unreliable components, v_r and v_u. In practice, a binary ‘mask’ $m(i,j)$ is used to indicate whether the acoustic evidence in each time-frequency region is reliable.

The reliable components v_r are directly available to the classifier in the form of a marginal distribution $f(v_r|\mathrm{C})$. Bounds may be placed on the values of the unreliable components, v_u, since their acoustic energies will lie between zero and the observed energy value, assuming additive noise. This approach is known as ‘bounded marginalisation’ [7].

3.5. Grouping by common azimuth

The binary mask required for missing feature speech recognition is derived from a model of auditory grouping by common azimuth. Additionally, an interaural level difference

(ILD) constraint is applied to channels above 2 kHz. The azimuths of the speech and noise, ϕ_s and ϕ_n, are derived from the pooled skeleton cross-correlogram. It is assumed that $\phi_s > \phi_n$ (i.e., that the speech lies to the right of the noise).

For channels below 2 kHz, the values in the mask are set according to,

$$m(i,j) = \begin{cases} 1 & \text{if } C(i,j,\phi_s) > C(i,j,\phi_n) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

In frequency regions above 2 kHz, grouping is performed as follows. First, $ild(i,j)$ in decibels is calculated for each channel i at time frame j. Then the ILD estimate is tested against the azimuth of the speech source, ϕ_S. The ILD is considered consistent with the azimuth estimate derived from ITD cues if,

$$\left| ild(i,j) - \Omega(i,\phi_s) \right| < 0.5\text{dB} \tag{3}$$

where $\Omega(i,\phi_s)$ is an ILD template precomputed in noiseless and anechoic conditions for channel i and azimuth ϕ_s. If this condition holds, then the corresponding mask value is set to one (otherwise zero). When the speech and noise are spatially separated, this heuristic effectively identifies the time-frequency regions that are dominated by the target speech source.

4. Evaluation and results

The model was evaluated on a 240 utterance subset of male speakers from the TiDigits connected digits corpus [9]. Auditory rate maps were generated from the training section of the corpus, and were used to train 12 word-level HMMs (a silence model, 'oh', 'zero' and '1' to '9'). All models were trained on clean signals.

The model was tested using a male utterance as an intrusion, taken from Cooke's corpus [8]. The amplitude of the intrusion was scaled to give signal-to-noise ratios (SNRs) of 0, 10, 20 and 200 dB. The intrusion and test utterance were convolved with left ear and right ear HRRIRs generated by Roomsim as described in Section 2.1. Using Roomsim the target speech source and interfering noise were placed at +20 and -20 degrees azimuth respectively. Reverberation conditions were controlled by selection of surface materials such that the reverberation times (RT_{60}) for the virtual room were 0, 0.3 and 0.6 sec. The theoretical limit of the missing data system was tested by forming masks that exploit 'a priori' information of clean speech regions. This is obtained by comparing the noisy speech sample and its clean counterpart during the mask generation and is a common way to demonstrate the limits of the missing feature approach with near optimal masks [7].

RT_{60} sec	Method	SNR 0 dB	SNR 10 dB	SNR 20 dB	SNR 200 dB
0	binaural	92.9	97.2	97.6	98.2
	a priori	96.3	97.2	97.6	98.2
0.3	binaural	61.6	89.2	94.1	95.0
	a priori	93.6	95.6	95.2	94.0
0.6	binaural	34.9	72.1	82.1	85.3
	a priori	92.4	95.1	95.5	94.5

Table 1: Speech recognition accuracy (percent) for continuous digit recognition in the presence of an interfering speaker and small-room reverberation.

Table 1 shows the experimental results as percentage recognition rate for 'a priori' masks and those derived from the binaural model. In the anechoic case, the binaural system almost reaches the performance using 'a priori' information. Clearly, increasing the reverberation time degrades the performance of the binaural system, but not the performance of the 'a priori' system. However, the performance of the binaural system shown in Table 1 remains substantially better than that of a conventional Mel frequency cepstral coefficient based ASR system [3].

5. Conclusion

The Roomsim program is an effective tool for researchers working in areas of acoustic signal processing, speech and hearing, who require to generate binaural audio signals sufficiently realistic for the assessment of signal processing algorithms aimed at improving the performance of e.g. speech recognisers, hands-free telephones and hearing aids. It also shows promise as an educational aid, of interest to those teaching in areas of acoustic signal processing, music technology, auditory perception, psychoacoustics, environmental acoustics, and digital signal processing.

The research results reported here suggest that the missing feature approach has considerable potential for robust ASR in reverberation, although there is room for improvement in the binaural mask estimation process.

6. Acknowledgements

The authors acknowledge the following supporting organisations: D.R. Campbell; Defeating Deafness (246:PAI:DC). K. Palomäki; EC TMR SPHEAR, The Academy of Finland (1277811). G.J. Brown; EPSRC GR/R47400/01. The authors gratefully acknowledge permission to include in Roomsim the HRTF data from CIPIC at the University of California [10] and the MIT Media Laboratory [11].

7. References

[1] Allen, J. B. and Berkley, D. A., "Image method for efficiently simulating small-room acoustics", *JASA* 65(4), 943-950, 1979.

[2] Lokki, T., "Physically-based auralization" PhD Thesis, Publications in Telecommunications Software and Multimedia. Helsinki University of Technology, 2002.

[3] Palomäki, K. J., Brown, G. J., and Wang, D. L., "A Binaural Processor for Missing Data Speech Recognition in the Presence of Noise and Small-Room Reverberation", Accepted for publication in *Speech Communication*, 2003.

[4] Peterson, P. M., "Simulating the response of multiple microphones to a single acoustic source in a reverberant room", *JASA* 80(5), 1527-1529, 1986.

[5] Gardner, W. G., and Martin, K. D. "HRTF measurements of a KEMAR dummy head microphone", In *Standards in Computer Generated Music*, Haus, G. and Pighi, I. eds., IEEE CS Tech. Com. on Computer Generated Music, 1996.

[6] Algazi, V. R. *et al*., "The CIPIC HRTF Database", IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2001.

[7] Cooke, M. P., Green, P. D., Josifovski, L. and Vizinho, A. "Robust Automatic Speech Recognition with Missing and Unreliable Acoustic Data", *Speech Communication*, 34, pp. 267-285, 2001.

[8] Cooke, M. P., "Modelling auditory processing and organization", Cambridge University Press, Cambridge, UK, 1993.

[9] Leonard, R. G., "A database for speaker-independent digit recognition", *Proc. ICASSP'84*, 111-114, 1984.

[10] http://interface.cipic.ucdavis.edu, March 2003.

[11] http://xenia.media.mit.edu/~kdm/hrtf.html, March 2003.