Judging the Speech Intelligibility of Large Rooms via Computerized Audible Simulations

3126 (C-8)

4E/C-8

Morten Jørgensen, Christopher B. Ickler, and Kenneth D. Jacob
Bose Corporation
Framingham, MA 01701, USA

Presented at
the 91st Convention
1991 October 4–8
New York

This preprint has been reproduced from the author's advance manuscript, without editing, corrections or consideration by the Review Board. The AES takes no responsibility for the contents.

Additional preprints may be obtained by sending request and remittance to the Audio Engineering Society, 60 East 42nd Street, New York, New York 10165, USA.

AN AUDIO ENGINEERING SOCIETY PREPRINT

Judging the Speech Intelligibility of Large Rooms via Computerized Audible Simulations

MORTEN JØRGENSEN, CHRISTOPHER B. ICKLER, AND KENNETH D. JACOB
Bose Corporation, Framingham, MA 01701

An audible simulation system has been developed which can simulate over headphones the sound field that a listener would hear from a sound system in a real room, based only on a computerized model of that system and room. This report describes the new system and the design of an experiment to test its ability to simulate the conditions responsible for speech intelligibility. The new simulation system is composed of three parts: a software program to predict the squared impulse response at a receiver location, software to convert the squared impulse response into binaural impulse responses, and hardware to convolve the binaural impulse responses with audio program material. The simulation system is capable of accounting for virtually all of the reflections and reverberation present in large rooms. The verification experiment, intended to begin the process of characterizing the accuracy of the audible simulation system, is partly complete. The objective of the experiment is to compare speech intelligibility scores obtained in a typical auditorium with those obtained using the new system simulating that same auditorium. Intermediate results show that subject-based speech intelligibility tests, conducted both in the auditorium and over the simulation system, yield scores with a high degree of precision over a wide variety of conditions for transmission of speech.

0 INTRODUCTION

Room acousticians, psychoacousticians, sound system designers and other audio professionals are concerned with the audible effect of changing something in the room in which they are working. For example, a concert hall designer may want to know the effect of changing wall materials, the sound system designer the effect of changing the orientation of a loudspeaker, and the psychoacoustician the effect of eliminating a reflection path. Usually, however, the cost and time required to make these experimental changes in rooms is prohibitive. In the case of an unbuilt room it may even be impossible. Instead, these investigators must rely on computer models or physical scale models to predict acoustical behavior, and from these predictions they must then judge the likely audible effect. Unfortunately, this process does not utilize the investigator's most powerful and direct tool for judging the audible effect of the proposed changes – his or her ears. A system that would allow the investigator to listen to the effect of a modification without working in the actual environment would be desirable.

The process of listening to such an audible simulation has been dubbed "auralization" by Kleiner [1]. In essence, *auralization is the act of listening to a signal which has been processed to sound like it would sound in a room, but without using the room itself in the audio signal path.* An auralization system is the hardware and software which permits auralization to occur. An example of a valid auralization system is a digital signal processing (DSP) system which filters program material in an attempt to simulate a room. In such a system the hardware can be instructed to change the way the signal is processed in order to approximate some physical change in the real room; the key is that the real room is not needed. On the other hand, a binaural-recording playback system is not an auralization system because the sound of the room was created by the room itself during the recording.

One of the most powerful applications of auralization is to the problem of designing sound systems in large rooms. With such a tool, sound system designers could obtain direct audible feedback of proposed sound systems in advance of installation. Furthermore, they could explore more options and correct only those defects which were aurally objectionable. They may even be able to guide the design of the room itself. Auralization would be an effective means of communicating the expected quality of a proposed sound system to the lay person, who is seldom versed in interpreting specifications of acoustical performance. In the end, auralization will improve the quality of sound systems in large rooms.

Perhaps the most urgent need for sound-system auralization systems is for simulation of speech intelligibility conditions. Auralization of speech, however, will need to be combined with a fast numerical prediction of speech intelligibility such as the speech transmission index (STI) [2]. The reason is that with auralization alone, the time and expertise necessary to obtain accurate intelligibility scores using human subjects would be prohibitive. It is the combination of these two methods of judging speech quality – auralization, and a fast, reliable, and accurate numerical algorithm – which we believe will lead to sound systems which perform as predicted, and which have better overall intelligibility.

Realizing these potential benefits requires much more than the development of an auralization system itself. An auralization system, although it may give the listener the general feeling that he or she is in a room, must first be proven to be accurate. A major distinction must be made between a plausible audible simulation and a realistic one, since in the former case only similarity is needed while in the latter quantitative proof of realism is required. For an auralization system to be an effective design tool, therefore, it must be proven scientifically to yield results which match those obtained from real rooms.

In addition to the new audible simulation system to be described here, there are others which may be capable of simulating sound systems in large spaces. The simulator used in the Archimedes project

[3] uses a large number of loudspeakers in an anechoic chamber. Applying program material to this system provides each loudspeaker with a signal having the correct characteristics for one or more of the reflected sound paths in the room being simulated. An advantage of this approach is that it maintains correct sound localization characteristics, even when the listener moves his or her head. The major disadvantage is that it requires an elaborate set-up in an anechoic chamber. The system is now in the verification phase, but only for the case of a small listening room.

Another auralization system uses a 1/10 scale model of the room, an omnidirectional source, and a scale-model head to measure an impulse response which is later convolved with various audio programs [4]. The disadvantages of this approach are that only omnidirectional sources can be used at present and that the cost, time, and space required for the model are relatively large.

A different system, now in development, allows a collection of discrete early reflections (along with any direct arrivals) to be convolved with program material through the use of PC-based DSP hardware [5]. Few details have been published and features are still being added to this system at the time of this writing.

Another approach [1] treats the entire room and sound system as a filter and uses extensive digital signal processing to supply filtered signals to the eardrums. The signals are presented to the listener via headsets or through two loudspeakers in an anechoic chamber. When presented through headsets, the approach has an advantage in that it does not require an anechoic chamber. The disadvantages are that the listener often localizes the sound inside his or her head, and that the approach requires expensive and difficult-to-obtain signal processing equipment. When the presentation is made through two loudspeakers, the advantage is that good localization can be achieved, but the disadvantages are that an anechoic chamber and the same signal processing equipment used in headset presentation are required.

Other auralization systems have been proposed which may be able to simulate sound systems in large spaces. They have been omitted only because they are similar to one of the approaches discussed above.

In this study a new audible simulation system is described which exploits the most recent advances in room modeling software and digital signal processing hardware. It consists of three parts. The first part is a room modeling program which calculates the sound propagation (in the model) from the input of the sound system to a receiver location within the room. The second part is an algorithm which converts this single-receiver response into two elaborate digital-filter descriptions which correspond to the sound transmission from the input of the sound system to an average listener's ears. The third part consists of the hardware required to process an arbitrary selection of program material – in real time – according to the two filter descriptions. The processed audio signal is played through headphones, ideally creating at the eardrums of the listener the same sound as they would hear if they were in the real room.

To begin the process of verifying the accuracy of this prototype auralization system, we have chosen to investigate first the crucial parameter of speech intelligibility. An experiment is described here which when completed will measure the accuracy of the new system in predicting speech intelligibility in a typical auditorium. The experiment is designed to allow comparison of intelligibility scores obtained using subjects taking a standard intelligibility test in the room to scores obtained when the same subjects take the same test while listening to the new simulation system.

1 AURALIZATION SYSTEM

As stated above, the prototype auralization system consists of three parts. Two of these generate the data which describe the simulated sound fields for each ear completely, regardless of program material. This is done before listening begins. In order to listen to a simulation, these descriptions are passed to the third part of the system, consisting of two powerful digital filters which process audio in real time according to these left-ear and right-ear responses. The output signals from these filters are converted to analog and sent directly to headphones for listening.

1.1 Approximating the Source-to-Receiver Impulse Response

The impulse response of a linear system (such as a sound system in a room) is the signal seen at the output of the system (a single location in the room) when a brief impulse or voltage spike is applied to the input (to the sound system, in this example). If the impulse response of the linear system is known, one can predict the output signal that will be produced from any input signal. This is because the input signal can be sampled, which reduces it to just a sequence of impulses of different sizes. For each impulse sent in, a proportionally sized impulse response comes out. The total output is just the sum of these many scaled, overlapping impulse responses. Using the impulse response in this way to compute the output signal from some known input signal is called convolution. In a very real sense, the impulse response is a complete description of the system. Measurement or prediction of the impulse response of a sound system in a room is all that is needed to determine the output at the receiver location from an arbitrary input signal. Finding or constructing that impulse response is fundamental to this and other auralization systems.

In this auralization system, a compact estimate of the square of the source-to-receiver impulse response is generated by an existing program used to predict the performance of sound systems in large rooms [6]. The predictions are based on an architectural model of the room and a loudspeaker model which includes full-space directional information. From the architectural description of the room and the location, orientation, and electrical power level applied to the speaker or speakers, the program makes an estimate of the major features of the squared source-to-receiver impulse response for any chosen position in the room. This approximation is called the Hybrid Energy Decay Curve (HEDC).

The program generates a unique HEDC for each octave band from 125 Hz to 4 kHz. An octave-band HEDC consists of two parts: 1) discrete early arrivals (including any direct arrivals), and 2) late statistical reverberation. A direct arrival is computed by adding source directional attenuation to the source-to-receiver inverse square loss. An image source method is used to predict the discrete early reflected arrivals. Each early arrival is represented by its octave-band strength in dB-SPL, its time of arrival, and its angle of incidence to the receiver location.

Prediction of discrete reflections having more than four intersections with room boundaries is beyond the power of the program. Furthermore, the assumptions of geometrical acoustics become less valid beyond third-order reflections, except in rare cases. Fortunately, the assumptions of classical statistical acoustics [7] become more valid at these later arrival times. For these reasons the second part of the HEDC consists of a continuous reverberant exponential tail in each of the six octave bands. In each band, the tail is defined by its

Sabine reverberation time and the total energy in the tail. The six tails all start at a common "splice time". With the direct sound and discrete early reflections, the tails form a complete estimate of the major features of the squared source-to-receiver impulse response.

1.2 Computing the Binaural Impulse Responses

The octave-band HEDC's are used as input to a signal processing program which converts them into an approximation of what would actually be measured at a listener's ears if he or she were present in the room when an impulse was applied to the sound system. These binaural impulse responses, therefore, are the source-to-left-ear and source-to-right-ear impulse responses. The binaural impulse responses are derived from the octave-band HEDC's and from measured responses of the head to plane waves from various directions, called head-related transfer functions [8,9]. The data for this last step was derived from a KEMAR manikin [10].

In the finished binaural impulse responses, each discrete early arrival has its own octave-band spectrum as specified in the six HEDC's. Each arrival is given interaural time delay and head shadow appropriate to its direction of incidence at the receiver location.

The reverberant tails of the binaural impulse responses simulate a random diffuse sound field that decays exponentially. Each octave band decays at a different rate as specified by the reverberation times in the HEDC's. The interaural cross-correlation between the two tails is correct for a diffuse field. The overall spectra of the tails are based on the total energy required by the HEDC's and on the diffuse field response of the KEMAR head.

The humidity-dependent effects of air absorption are included in both the discrete and reverberant parts of the binaural impulse responses. The responses also include equalization to compensate for the headsets used. The resulting binaural impulse responses are each 65,536 points long and are sampled at 48 kHz, giving a length in time of 1.37 seconds. Each point is stored as a 24-bit integer.

These first two parts of the auralization system – estimation of the source-to-receiver squared impulse response and conversion of this response to binaural impulse responses – complete the modeled description of the sound field that would be measured at the listener's ears when the sound system is provided with an impulse at its input. These two parts are not active during the auralization, or listening process.

1.3 Real-Time Digital Processing

The last step is where digital audio equipment convolves the binaural impulse responses with program material (speech for example) to simulate how it would sound in the real room. In general, the convolution takes place either as a discrete convolution in the time domain, or as an equivalent filtering in the frequency domain.

The heart of the convolution process is the digital filter that is programmed to mimic the room. Its function is conceptually simple – to continuously convolve the modeled binaural impulse responses with digital audio program material in real time. Since these impulse responses are more than 64,000 samples long, this is beyond the capability of any general-purpose DSP system now available. Instead, the simulator uses two Austek A95001 Frequency Domain Processors [11]. As delivered, an A95001 is capable of convolving stereo audio with two impulse responses 32K samples long. This is only 0.7 seconds of reverberation, so that if the simulated room were to have a reverberation time of 3.5 seconds each sound would go silent after only 12 dB of decay. With modified firmware from Lake DSP [12], an A95001 can convolve mono digital audio with one impulse response of length 128K (131,072 samples). Such impulse responses are 2.7 seconds long. Two modified A95001's were used in the simulator, one for each ear. The simulation system can be reprogrammed for a different simulation in less than one minute.

2 EXPERIMENT TO TEST AURALIZATION SYSTEM ACCURACY

The strategy used in this experiment is to measure speech intelligibility in several positions in a real room using subjects and a standard speech intelligibility test, then to repeat the test using audible simulations of that room. The room, loudspeaker setup, and listener positions are chosen so as to create regions of good and bad intelligibility. Thus if the room and the simulator were to give the same results, the simulator would be accurate in portraying speech intelligibility conditions for that room.

Because the auralization system is intended for listening, it must be tested by listening. Therefore, if speech intelligibility is to be measured, a listening-based intelligibility test is required. Fortunately, controlled subject-based measurements of intelligibility can yield precise results. Even small differences in intelligibility between the room and the simulation can be resolved with such a technique. This form of subject-based testing is also the most direct way to estimate the intelligibility of the sound system in actual use. Virtually all other means of estimating intelligibility were developed to predict subject-based scores.

2.1 Standardized Intelligibility Test

A standardized intelligibility test, as specified by the American National Standards Institute (ANSI) [13], was chosen, with some exceptions listed below. The stimuli used in this standardized test are standard monosyllabic, phonetically balanced English words. There are twenty lists of fifty words each.

In preparation for this experiment, all 1000 test words were recorded embedded in the carrier sentence: "Would you write *word* now." These phrases were recorded with a sample rate of 44.1 kHz and 16 bits per sample. The recordings were made in an anechoic chamber, using a B&K 1/2 inch omnidirectional, instrumentation grade, free-field microphone, and a talker-to-microphone distance of one-half meter. A signal-to-noise ratio of at least 30 dB was achieved in order to prevent background noise from affecting the intelligibility of the recordings [14]. Only one talker was used rather than the five required by the ANSI standard.

Only eighteen of the twenty word lists in the ANSI standard were used in this study. Statistical analysis of earlier work [15] revealed that, for identical experimental conditions, word lists 14 and 15 each gave 4-5% lower speech intelligibility than the other lists. The small number of positions in this experiment made it essential that all the word lists should have the same degree of difficulty; thus lists 14 and 15 were not used.

2.2 Auditorium and Auditorium Model

A typical auditorium was chosen. Nevins Hall is a multi-purpose auditorium in the Boston metropolitan area, and is used for town meetings, voting and other functions. The auditorium is approximately 35 meters long by 27 meters wide by 12 meters high and has a total volume of 9,500 cubic meters. Walls and ceilings are made of gypsum and the floor is wood parquet. The rear wall, ceiling, and the second floor balcony are curved, making it an acoustically complex space. The curved rear wall was suspected of causing focusing effects on the main floor and widely varying acoustic conditions within relatively short distances. The measured reverberation time was 3.2 seconds averaged over the 1-, 2- and 4- kHz octave bands at 60% relative humidity.

A computerized architectural model of the auditorium was made by drawing N-sided planes (N<10) and assigning each surface with a material chosen from a standard list. (Materials are described by their octave-band Sabine diffuse-field absorption coefficients, as given by the manufacturer or standard acoustics texts.) A simplified version of the model is shown in Fig. 1. Plan dimensions were derived from blueprints, while vertical dimensions were measured in the room. The curved surfaces were approximated by multiple planes. It should be noted that the absorption coefficients of gypsum in the material list were slightly changed to match predicted reverberation times with measured times. It is a good idea to do this in modeling rooms that already exist, since the actual construction may not match the standard absorption data.

2.3 Sound Source and Sound Source Model

The sound source was an Electro-Voice HP6040A constant-directivity horn with a DH2 compression driver, crossed over at 800 Hz to an Electro-Voice TL806AX bass-to-midrange speaker. The crossover was an Electro-Voice XEQ 808 passive network. The directivity of the horn was Q=21 averaged over the 1-, 2- and 4- kHz octave bands. The horn was placed on a speaker stand on the stage, 2.8 meters above floor level and aimed off-center as shown in Fig. 2. The bass-to-midrange speaker was positioned beneath the horn and aimed in the same direction. Off-center aiming was chosen to ensure that some positions of poor speech intelligibility were likely to be found for the test.

The speaker was equalized with a one-third octave band equalizer (Rane GE27) to have a flat spectrum at eight meters on axis in the room from 125 Hz to 2 kHz, and a roll-off of -3 dB per octave at higher frequencies. Deviations from this did not exceed 1 dB. Before any testing was conducted the sound system was swept with a sine-wave at a loud level to find and eliminate any audible buzzes or distortion. The sound level at this same location was adjusted to 79 dB-A.

Sound sources were represented in the computer room modeling program by their full-space directional characteristics and sensitivity in the octave bands from 125 Hz to 4 kHz, combined with their electrical power inputs. They were modeled by describing their location in the computerized room and their aiming angles.

2.4 Listener Positions

Six listener positions were chosen with the intent of obtaining speech intelligibility scores covering the range from 60% to 95%. The sound source, room, and listener positions were selected to create conditions where two factors were responsible for reduced intelligibility: 1) too much reverberation, or 2) poor high frequency coverage.

Positions were selected using the following procedure. First, the speech test was reproduced over the loudspeaker system, using the same setup that would be used later with the subjects. With the speech test playing, the experimenters listened at a number of different locations throughout the hall in order to make rough estimates of the actual intelligibility. When this rough mapping of intelligibility was completed, a position of excellent intelligibility on the main axis of the horn, and a particularly poor position off axis were chosen; these locations became listener positions 1 and 6 respectively. Then a position was chosen which seemed to be about halfway (in intelligibility, not distance) between positions 1 and 6. This led to the determination of position 3. Then a position which was noticeably worse than position 3 was found, leading to the choice of position 4. Then the process of determining a position halfway between two others was repeated for positions 1 and 3, and positions 4 and 6, leading to the choice of positions 2 and 4. This method was surprisingly easy to carry out, and all six positions were chosen within an hour and with reasonable confidence that a good spread of intelligibility conditions would be found.

The locations of these positions are shown in Fig. 2. Listener position 1 was on axis of the horn, position 2 in the good coverage area of the horn, while positions 3, 4, 5 and 6 were in the fair to bad coverage area of the horn. Listener positions corresponding to those used in the real room were located in the computerized room model.

2.5 Listeners

Five female listeners in the age range from 23 to 50 years old were selected. All had a college education and American English as their native language. Their hearing thresholds were measured using an audiometer meeting ANSI S3.6-1989 [16]. (Note that this is a different ANSI standard, referring to audiometric tests. All other references to ANSI testing in this report refer to the intelligibility testing standard ANSI S3.2-1989.) All subjects had hearing thresholds below +15 dB from 125 Hz to 8 kHz.

2.6 Design of Subject-Based Testing

Listeners were trained in advance of the actual testing. First, they listened to dry recordings of the word lists for three hours (with rest breaks), writing down the words they perceived. On the following day they were trained using the test procedure in the real room for three hours. On the third day they trained using headsets and the auralization system for three hours. At the end of the training period, all listeners had heard all of the ANSI words at least three times. The actual testing was begun when the listeners were comfortable with all procedures, and when their scores in the different test situations had reached a stable level of performance. The listeners were not told that the first sessions were for training purposes only, nor did they know when actual testing started. They were not informed of the purpose of the experiment or of any relationship between the in-room listening and the headset listening (auralization).

The number of ANSI words tested in each listener position was based on the desired worst case error of the speech intelligibility scores in the listener positions. The error of these scores is defined as half the width of the confidence interval. Assuming normally distributed data, the confidence interval is given by:

$$t(N-1)_{1-\alpha/2}\frac{\sigma_{\mathrm{data}}}{\sqrt{N}} \tag{1}$$

where,

N is the number of scores for this position,

t(N-1) is Student's t-distribution with N-1 degrees of freedom,

α is the significance level (5%), and

σ_{data} is the standard deviation of the scores.

A worst-case standard deviation of 13%, as obtained independently by Jacob [17], and Jørgensen and Petersen [18], was expected. In this experiment it was decided to use no less than 2,750 words per combination of room, sound source and listener position. Using Eq. 1 above, this gives an expected worst-case error of 3.5%. (It should be noted that this is not as many words as prescribed by the ANSI standard, but the expected worst-case accuracy was acceptable for this experiment.)

The order of experimentation was randomized to average out the effects of variables such as listeners, word lists, and day-to-day variations which could not be controlled. Restrictions in the ability to randomize made the two subject-based tests slightly different, as described in the sections that follow.

2.7 Limitations of Experiment

The experiment described above was performed. Actual subject-based intelligibility scores were obtained for both the in-room and auralization system conditions. After these scores were obtained, however, an error was found in the way that the computerized model was made, invalidating the comparison of scores obtained from the auralization system with those obtained from the room. The error consisted of an incorrect accounting for of the effect of obstructing room planes (such as balconies) in the computer model. The result was that the predicted squared impulse responses (the HEDC's), while containing the correct prediction of the direct arrivals, discrete arrivals, and the diffuse reverberation, contained extra discrete early arrivals which should not have reached the receiver because of obstructing planes. For this reason, the experiment will need to be performed again with the correct predicted impulse responses. Nevertheless, the subject-based testing which has been completed contains valuable information which can guide us in our repetition of the experiment, and this information is presented in the following sections.

3 SPEECH INTELLIGIBILITY TESTING IN THE REAL ROOM

A variety of speech intelligibility conditions, ranging from excellent to poor, over the listener positions was desired. Results of the subject-based speech intelligibility tests conducted in the auditorium at the proposed listener positions would either confirm our choice of listener positions or indicate which positions should be changed. At a future date, these in-room tests will be repeated in order to obtain final results for this experiment. This preliminary subject-based test also allowed us to gain experience in the experimental design proposed for the experiment, and to determine with confidence the number of words required to obtain good subject-based data.

3.1 Equipment

The equipment for the subject-based tests in the auditorium is shown in Fig. 3. Speech stimuli for the test were stored digitally on a hard disk connected to an Apple Macintosh computer. The Macintosh converted the digitized speech to analog and sent it through a 100 Hz high-pass filter to remove any audible low frequency hum or noise. A pre-amplifier, one-third octave band equalizer, power amplifier and passive cross-over network completed the signal path to the two-way loudspeaker system located in the auditorium. Five subjects were distributed among six listening positions in the room.

3.2 Procedure

The subject-based test was administered as described in section 2.1. The speech presentation was controlled by a HyperCard program running on a Macintosh IIcx computer. The program controlled the presentation rate (15 phrases per minute) and the order of the phrases. Each time a word list was used, the program presented the target words in a different random order to reduce recognition effects. The number of words presented in each listener position was no less than 2,750.

The testing was conducted in eleven sessions, each consisting of six blocks. In a block the five listeners were distributed randomly among the six listener positions. (In the sixth position, binaural recordings were made at a KEMAR manikin's ears for a future experiment.) Blocks within a session differed by 1) a different random assignment of listeners to listener positions, and 2) a different word list. In a session each listener heard one word list at each one of the six listener positions.

Sessions differed from each other in that 1) they each used a different set of six word lists, 2) the order of the words in any list was never twice the same, and 3) the order in which a subject visited the six positions was always different (if a listener was tested in the order 1-2-6-3-5-4 in one session, the order in the next session might have been 2-3-1-4-6-5).

A session was designed as an incomplete, balanced block design [19], with listener position, listener, and word list as the main variables. Sessions were conducted in six blocks, thereby confounding the word list effect (a main effect) with the block effect, or in other words, if a significant effect from block to block was found, it was impossible to say whether it was due to 1) a different degree of difficulty of the word lists or 2) a difference in the experimental conditions in the two blocks.

Sessions were conducted in groups of two with a break of 10 minutes between the sessions. Four sessions (two groups) were conducted every day with a break of one hour between the two groups. After each session, analysis of variance (ANOVA) was made to test for significant main effects on a 5% significance level. A significant listener position effect was found in all eleven sessions, showing the variation in speech intelligibility at the different listener positions. In five sessions either a significant listener or word list effect was found. These effects were not considered serious because they were caused by different word lists and different listeners.

3.3 Data Reductlon

For each word list presented, five out of six listener positions were occupied by a subject. When one word list was complete the five word sheets were scored, thereby yielding one intelligibility percentage score for each of the five listener positions. The results of all the presentations for a given position are given as a mean percentage of correctly understood words with a 95% confidence interval.

Assuming normally distributed data, the confidence interval of the mean scores was computed using Eq. 1.

3.4 Results

The results of the preliminary in-room intelligibility testing are shown in Fig. 4. From the results it can be seen that the goal of obtaining a wide variety of scores was achieved. In fact, the scores can be seen to decrease monotonically from listener position 1 (95%) to position 6 (58%). Secondly, the confidence intervals computed for each mean score are low (<±3.1%), which shows that the mean intelligibility scores are precise. These results allow us to conclude that the choice of listener positions needs no adjustment, and that the amount of subject-based testing could be reduced by using fewer words. These results also indicate that the simple procedure used to make our initial choice of listener positions, based on listening in the room to the speech test, is a good one.

4 SPEECH INTELLIGIBILITY TESTING USING THE AURALIZATION SYSTEM

The same subject-based speech intelligibility test used in the real room was used to obtain speech intelligibility scores using the auralization system. An incomplete computer model of the room was used to generate Hybrid Energy Decay Curves in listener positions corresponding to those used in the real room. Then the octave-band HEDC's were converted to binaural impulses which were downloaded to the audio equipment performing the real-time convolutions with the speech-test program material. The listeners heard the test words via headsets, as a simulation, and wrote down the perceived words.

4.1 Apparatus

The set-up for the subject-based testing using the auralization system is shown in Fig. 5. The speech program material from the Macintosh computer was fed to a DAT recorder which simply performed an analog-to-digital conversion. In advance, an IBM PC had stored all of the binaural impulse responses on its hard disk for use as needed. To simulate a listening position, the two binaural impulse responses for that location were downloaded from the PC to the Austek A95001 filters. From then on, the A95001's continuously convolved the incoming digital speech with the binaural impulse response coefficients. The digital audio outputs of the A95001's were converted to analog and presented via headsets to the listeners who were located in a separate listening room.

4.2 Procedure

The procedure to measure intelligibility using the auralization system included room modeling, computation of the octave-band HEDC's, computation of the binaural impulse responses, convolution, and subject-based testing. Each of these steps is described in detail below.

4.2.1 HEDC's

From the model of the room, six octave-band HEDC's were computed for all six listener positions, as described in section 1.1. The HEDC's used were not correct for the six listener positions due to an error in configuring the modeling program (see section 2.7); the error resulted in the prediction of extra discrete reflections which should have been blocked by obstructing room planes. HEDC's for this experiment contained from 65 to 76 discrete early arrivals, including first, second, and third-order reflections. The splice times (when the exponentially decaying reverberant tail was applied) ranged from 205 to 234 milliseconds after the beginnings of the HEDC's. Computation of the HEDC's for each listener position took about 15 minutes on a Macintosh IIcx.

4.2.2 Binaural Impulse Responses

Once the HEDC's for a given listener location were computed, they were passed from the room modeling program to the signal processing program for conversion to binaural impulse responses as described in section 1.2. Construction of a pair of impulse responses for a single listening position required about 30 minutes on a Prime 9750 minicomputer. For this experiment modeled binaural impulse responses for all six listening positions were computed several days in advance of the testing.

The humidity used in the calculation was 60% relative humidity. This value is within 10% of the humidity measured in the auditorium during the listening tests described in section 3.

4.2.3 Convolution

With the modified firmware from Lake DSP, an A95001 was able to convolve mono digital audio with one impulse response of length 128K (131,072 samples) as described in section 1.3. The signal processing program described above was not ready for this new length. The 64K point modeled binaural impulse responses were therefore padded out with an additional 64K of zeros before use. Such impulse responses are 1.37 seconds long at a sampling rate of 48 kHz. The octave-band reverberation times were less than 3.5 seconds. This means that any simulated sound from the room decayed at least 23 dB before the end of the filter.

4.2.4 Subject-Based Testing

The subject-based testing of the simulator was performed as described by the ANSI standard, under the same experimental conditions as mentioned in section 3.2. The sound level was adjusted to the same level (79 dB-A) as the in-room testing.

The subject-based testing using the auralization system was conducted in twelve sessions, each defined by six blocks. In a block, the subjects listened over headsets to the simulations of one listener position, writing down the perceived words from one word list. In each session every one of the listeners heard one word list from each one of the six simulated listener positions. After each block the listeners changed headsets. Sessions deviated from each other by 1) a different and random order in the simulated listener positions and 2) different word lists.

A session was designed as an incomplete, balanced block design with simulated listener position, listener, word list and headset as the main variables. By conducting the experiment in blocks as described above, the simulated listener position effect was confounded with the block effect. This is unfortunate, since the simulated listener position effect is the most important single main effect. It was decided to run each block in two parts, first for two listeners and later on the same day for the other three listeners. Thus the simulated listener position effect only was partly confounded with the block and the word list effect within a session. The confounding was necessary due to both time constraints and only having one auralization system. The detrimental effect of this confounding was decreased because the design essentially is twelve replications of the same session. Major efforts were made to make the experimental conditions the same from block to block, thereby decreasing any block effect.

The scores obtained using the auralization system are based on 3,000 words for each position. Data were reduced in the same way as described in section 3.3.

4.3 Results

The results of the preliminary subject-based testing using the auralization system are shown in Fig. 6. The confidence intervals are low and comparable to those obtained in the in-room testing. This suggests that the listening conditions and headsets used for auralization do not introduce any special effects which lead to more variance in subject-based scores. Subjects are equally adept and consistent in listening to this auralization system as they are in the real room. This result is important in that it suggests that the same precision can be obtained using the auralization system as with in-room testing under the same experimental conditions. The results also show that the confidence intervals are significantly lower than the desired worst-case interval predicted in section 2.6, indicating that the time required to conduct the test could be reduced by using fewer words.

5 CONCLUSION

A prototype auralization system has been described and an initial experiment designed to test its accuracy as a predictor of speech intelligibility. The auralization system, which is capable of simulating a sound system in a large room without requiring the room itself, attempts to overcome disadvantages in other approaches. First, it depends on a sound system-and-room modeling program which has been shown to accurately predict speech intelligibility [20]. Second, the algorithm used to construct binaural impulse responses accounts for the spectrum of each discrete direct and reflected arrival, the spectrum and frequency-specific decay rates of later-arriving diffuse reverberation, the effect of air absorption, the effect of the human head, and the effect of headset playback. The result is a relatively comprehensive treatment of the major factors affecting how sound arrives at a listener's eardrums in a real room. Third, the system performs its processing of an arbitrary selection of high fidelity program material in real time. Fourth, the system processes with binaural responses which are up to 131,072 points long, allowing it to simulate even very reverberant spaces. The system is therefore theoretically capable of simulating the sound of most sound systems in most environments. And fifth, the system is headphone-based and can be used in any quiet environment; it does not use an elaborate setup of loudspeakers in an anechoic chamber.

This prototype auralization system is undergoing extensive testing to determine the accuracy with which it can simulate real sound systems in real rooms. Without such verification this, or any auralization system, can at best be described as giving plausible, but not necessarily realistic, simulations. As part of this testing, an experiment was designed and begun to investigate the system's ability to predict intelligibility in a typical room with a sound system, where the intelligibility ranged from excellent to poor.

Preliminary subject-based testing was conducted both in the test auditorium and using the new auralization system. The focus of the testing was to: 1) determine the level of precision in measuring speech intelligibility which could be obtained under these proposed experimental conditions, 2) gain experience in executing this experimental design, and 3) confirm that the listener positions chosen yielded intelligibility scores which ranged from excellent to poor. The results show that, using the experimental design and under the experimental conditions described here, high precision can be obtained in the subject-based scores; in fact, the data indicate that significantly fewer words could be used while maintaining the desired precision. This result means that the time needed to conduct subject-based testing can be reduced. The results also showed that a simple method of determining listener positions based on auditioning the speech test led to the intended variation of actual subject-based scores across the six listener positions.

Now that the preliminary part of this experiment is complete, testing in the room and on the auralization system using the correct predictions of the source-to-receiver squared impulse response can proceed. This test will allow the direct comparison of scores obtained from in-room testing with those obtained using the new auralization system. In addition, it is anticipated that the speech transmission index will be both measured in the room, and predicted in the computer room modeling program, so that the STI's prediction of speech intelligibility scores can be compared to the actual in-room scores.

Just as the quality of architectural designs are improved by proper use of scale modeling and image rendering – both of which provide feedback to the architect's primary sense for judging quality (the visual sense) – we can look forward to better acoustical designs through the use of auralization.

Recent developments in the area of computer modeling and DSP hardware ensure that the quality of auralization systems will improve and probably become less expensive. Before auralization systems become part of the sound system designer's toolkit, however, their accuracy must be characterized, and this study begins this process. As these verification studies proceed, and the accuracy of auralization is improved, it is our hope that auralization will allow the designer to use his or her primary sense – hearing – to judge the quality of their unbuilt designs.

The authors would like to acknowledge Edwin Johnson and Joseph Kutil, members of the acoustic research staff, for their technical support and contributions to this study.

10 REFERENCES

[1] M. Kleiner, P. Svensson, and B. Dalenbäck, "Auralization: Experiments in Acoustical CAD," presented at the 89th AES Convention (Los Angeles, 1990), preprint #2990.

[2] H. Steeneken and T. Houtgast, "A Review of the MTF Concept in Room Acoustics and Its Use for Estimating Speech Intelligibility in Auditoria," *J. Acoust. Soc. Am.*, vol. 67, no. 1 (1985).

[3] S. Bech, "Electroacoustic Simulation of Listening Room Acoustics; Psychoacoustic Design Criteria," presented at the 89th AES Convention (Los Angeles, 1990), preprint #2989.

[4] K. Oguchi, S. Ikeda, and M. Nagata, "Application of Computer Simulation and Scale Model Testing to Room Acoustical Design," presented at the 89th AES Convention (Los Angeles, 1990), preprint #2991.

[5] S. Berkow, private correspondence.

[6] Modeler Design Program v3.1. Modeler is a registered trademark of Bose Corporation.

[7] L. Beranek, Acoustics, (reprinted by the Am. Inst. Phys. for the Acoust. Soc. Am., New York, 1986), part XXIV.

[8] J. Blauert, Spatial Hearing, MIT Press, Cambridge, USA (1983).

[9] E. Cohen, "Technologies for Three Dimensional Sound Presentation and Issues in Subjective Evaluation of the Spatial Image," presented at the 89th AES Convention (Los Angeles, 1990), preprint #2943.

[10] Manikin Measurements, Proceedings of a Conference Organized by M. D. Burkhard, Industrial Research Products, Illinois (1978).

[11] P. Single and D. McGrath, "Implementation of a 32768-Tap FIR Filter Using Real Time Fast Convolution," presented at the 87th AES Convention (New York, 1989), preprint #2830.

[12] Lake DSP Pty. Ltd., 4/15 Baden Street, Coogee 2034, Australia.

[13] ANSI S3.2-1989, "Method For Measuring The Intelligibility Of Speech Over Communication Systems," American National Standards Institute, New York, 1989.

[14] K. D. Kryter, "Methods for the Calculation and use of the Articulation Index," *J. Acoust. Soc. Am.*, vol. 34, pp. 1689-1687, (1962).

[15] K. Jacob, "Correlation of Speech Intelligibility Tests in Reverberant Rooms with Three Predictive Algorithms," *J. Audio Eng. Soc.*, vol 37, pp 1020-1030 (Dec., 1989).

[16] ANSI S3.6-1989, "American National Standard Specification for Audiometers," American National Standards Institute, New York, 1989.

[17] K. Jacob, "Correlation of Speech Intelligibility Tests in Reverberant Rooms with Three Predictive Algorithms," *J. Audio Eng. Soc.*, vol. 37, pp 1020-1030 (Dec., 1989).

[18] M. Jørgensen and J. Petersen, "Vurdering af Taleforståelighed fra PA-anlœg," ("Judging Speech Intelligibility from PA-systems"), The Acoustics Laboratory, Technical University of Denmark (1990) (in Danish).

[19] C. R. Hicks, "Fundamental Concepts in the Design of Experiments," third ed., CBS College Publishing, New York (1982).

[20] K. Jacob, T. Birkle, and C. Ickler, "Accurate Prediction of Speech Intelligibility without the use of In-Room Measurements," *J. Audio Eng. Soc.*, vol. 39, No. 4, pp 232-242 (Apr., 1991).

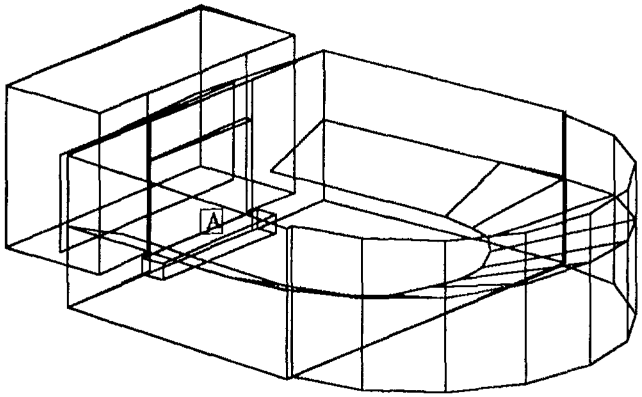

Fig. 1. A simplified model of Nevins Hall, Framingham, Massachusetts is shown. The room is used for town meetings and other events. Its mid-band reverberation time is 3.2 seconds. The speaker location used in this study is indicated by A.

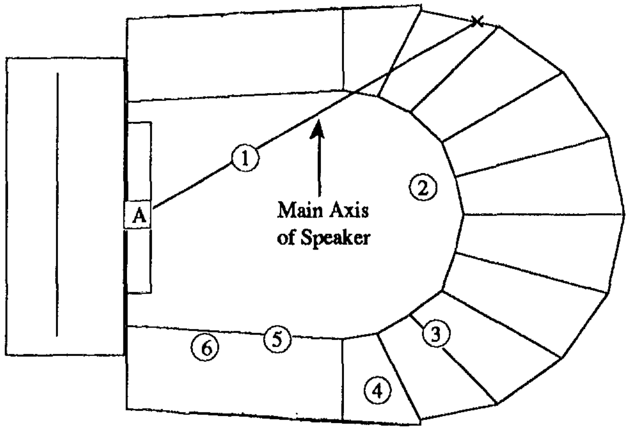

Fig. 2. The six listener positions used in this study are indicated by their numbers (1-6). The sound source (indicated by A) is located on a speaker stand 2.8 meters above the floor, and is aimed as shown. Position 1 is on-axis, 8 meters (0°, 8m) from the sound source. The location of the other positions are: pos. 2 (25°,19m), pos. 3 (52°, 21m), pos. 4 (66°, 12m), pos. 5 (72°, 13m) and pos. 6 (90°,10m). All positions are on floor level.

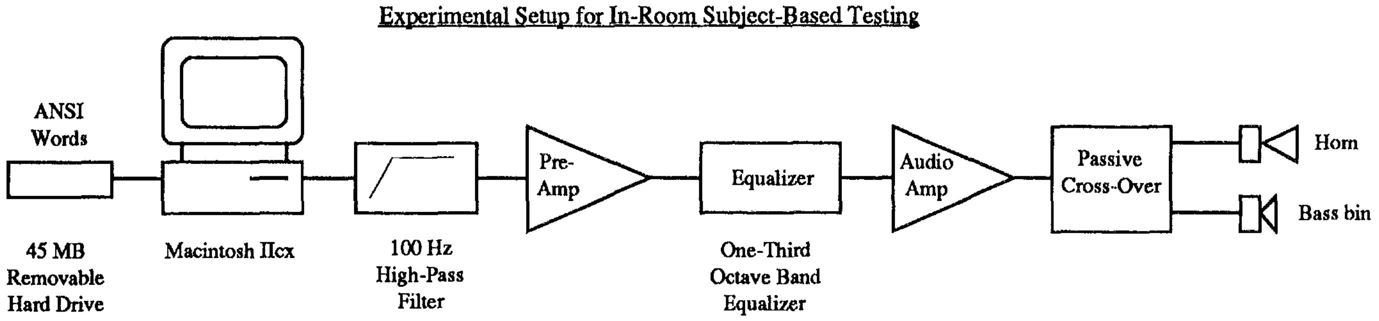
Experimental Setup for In-Room Subject-Based Testing
ANSI Words
45 MB Removable Hard Drive
Macintosh IIcx
100 Hz High-Pass Filter
Pre-Amp
Equalizer
One-Third Octave Band Equalizer
Audio Amp
Passive Cross-Over
Horn
Bass bin

Fig. 3. The setup for the in-room subject-based testing is shown. The digitized ANSI words were stored on a removable hard drive, which was connected to a Macintosh IIcx computer. From the D/A converter at the output of the Macintosh, the program material was sent to a 100 Hz high pass filter to reduce audible low frequency noise and hum. A pre-amplifier, equalizer, amplifier and passive cross-over filter completed the signal path before the signal was sent to the EV-6040A horn and the EV-TL806AX bass-to-midrange speaker.

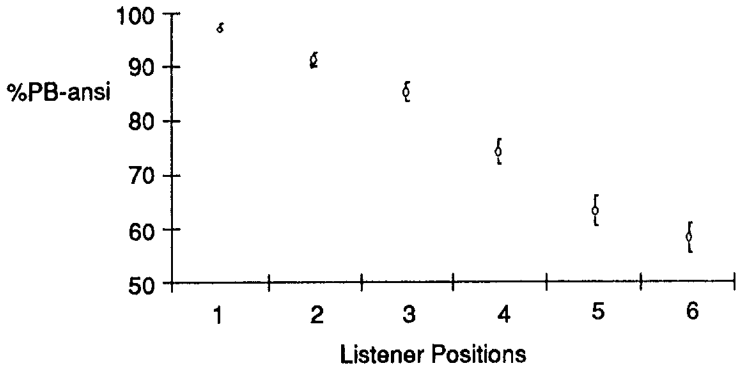

Fig. 4. Mean scores (open circles) obtained from speech intelligibility testing in the room are shown, along with 95% confidence intervals (vertical bars). Data shows excellent spread of intelligibility scores, ranging from about 60 - 95% speech intelligibility. Data also shows high precision (low confidence intervals).

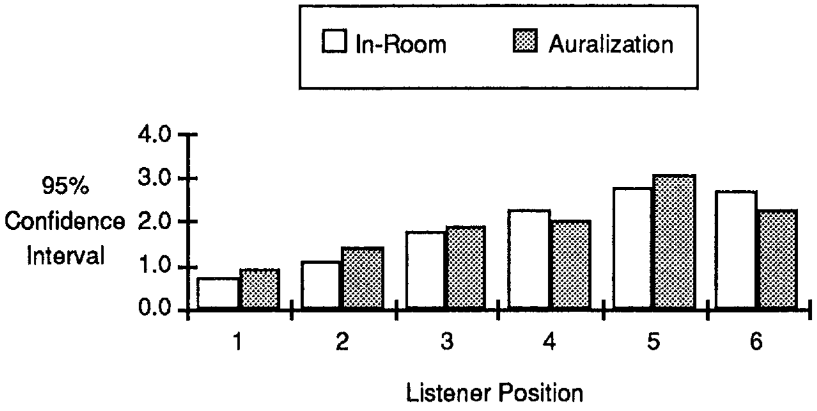

Fig. 6. 95% confidence intervals for mean speech intelligibility scores obtained from the actual room, and using the auralization system are shown. The data show that the same precision can be obtained using the auralization system as can be obtained in the room. The data also show that the precision is higher than expected, meaning that the number of words could be reduced in future experiments.

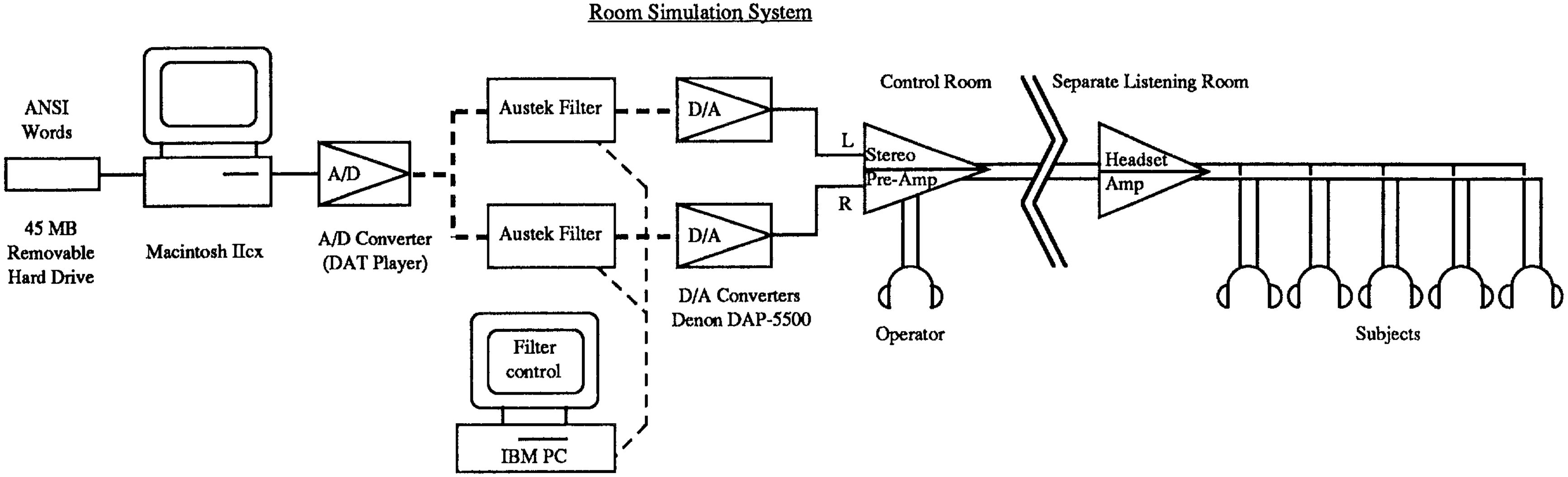

Fig. 5. The setup used for subject-based testing using the auralization system is shown. The Austek filters receive two files representing the binaural impulse responses from the IBM PC. They then continuously convolve the two responses with the ANSI test phrases coming from the Macintosh computer. The processed audio is converted to analog and presented to the listeners via headsets.