**Using Subject-Based Testing to Evaluate the Accuracy of an Audible Simulation System** 3725 (B2-AM-2)

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

**Presented at the 95th Convention 1993 October 7–10 New York**


# AES



## AN AUDIO ENGINEERING SOCIETY PREPRINT

# Using Subject-Based Testing to Evaluate the Accuracy of an Audible Simulation System

**MORTEN JØRGENSEN, CHRISTOPHER B. ICKLER, AND KENNETH D. JACOB**

*Bose Corporation, Framingham, MA 01701*

Audible simulation systems will allow sound system designers and their clients to judge the quality of a sound system before installation based only on the sound of a computerized model. However, if developers of audible simulation systems do not determine the accuracy of their simulators before they are used to design real sound systems, then some users may be badly misled. Without knowing the accuracy of the simulations, users simply will not know if the sound of the simulated system will have any useful similarity to the sound of the system actually installed. This paper addresses the major issues relating to the problem of determining simulator accuracy. A new term for this field – *authentication* – is used to describe the scientific process of quantifying to what extent people hear the same thing in the simulated environment as they hear in the real environment. At the core of authentication experiments are subject-based listening tests, where listeners' responses to the simulations are compared to their responses when they listen to the actual sound system in its environment. The output of an authentication experiment is three quantities related to the accuracy of the simulations. With this information about simulator accuracy to guide them, sound system designers and their clients will be able to confidently use an audible simulation system to listen to a proposed design.

## 0 INTRODUCTION

Sound system designers, room acousticians, concert hall designers and other audio professionals frequently use computer modeling tools to predict the behavior of sound in spaces such as auditoriums, conference rooms and concert halls. The fundamental advantage of using computer modeling tools is that without actually being in the physical environment, it is possible to predict how the sound will behave in that environment. Just as important, changes can be made to the computer model to find out what will happen to the sound, thereby avoiding making expensive changes to the actual physical environment.

Existing computer modeling tools present their results in the form of graphs, figures, maps, and tables of numbers. In every case, the output is in numeric form. However, in the last few years a number of research efforts have been devoted to the development and use of audible simulation technology, which promises to allow computer modeling tools to produce output to

which we can directly listen. To grasp the profound benefit of such a technology, one has to think about how the quality of computerized models are currently judged. Basically, we have to *imagine* the sound quality of a proposed sound system based only on numerical predictions. For example, we must imagine the loudness based on predictions of sound pressure levels, and imagine the speech intelligibility based on predictions of the speech transmission index (STI).

What we would like to do, of course, is listen to the computerized models, so that we could judge the sound quality directly. And that is exactly what audible simulation technology promises to allow us to do. Audible simulation systems will allow designers and their clients to audition a proposed system, to use the best tool they have for judging sound quality – their ears – while listening only to a computerized model of a sound system in a room.

The potential impact of this technology is huge. First, sound system designers will be able to design better sound systems because they will be able to hear their progress as they design. And just as important, they will be able to communicate better the quality of a design to their customers. Instead of asking customers to evaluate the sound quality of the proposed system by looking at numerical predictions, they will be able simply to ask their customers whether they like (or do not like) what they hear.

Before we can realize these benefits, however, audible simulation systems must be proven both useful and trustworthy to their users. We mean useful in the sense that designers ought to be able to work on a simulated sound system the same way they work on an actual system. For example, designers use equalization to change the tonal balance of an actual system. They ought therefore to be able to apply equalization to the simulated system. But, this alone does not guarantee that the effect of that change will be correct. The audible effect of changing a factor on the simulated sound system may or may not result in the same audible effect as when the factor is changed on the actual

system. That is only true if the simulations are accurate. Therefore, users also need to know about the simulator accuracy. It is only with this knowledge about simulator accuracy that users will consider an audible simulation system trustworthy.

The importance of quantitatively determining the accuracy of audible simulation systems can not be underestimated. Designers and their clients can easily be misled if they do not know to what extent the simulations bear resemblance to what the actual sound system will sound like. For example, if the simulator creates higher speech intelligibility than the real system because the simulator can not include the effects of background noise, both the designer and the client could be seriously disappointed in the real system.

There are a number of research and development efforts underway in audible simulation technology. Some research groups are developing audible simulation systems aimed primarily at concert hall design [1, 2]. Others are developing simulation systems to simulate loudspeakers in listening rooms [3]. Again others, like we, are involved in developing simulators to aid in sound system design [4]. However, to our knowledge there is no reported scientific evidence by anyone (including us) that the simulation systems under development bear any resemblance to reality.

We believe that it is possible to determine the accuracy of a simulation system in a scientific manner. And this paper proposes a strategy – called *authentication* – to do that. (It is important here to stress that this is our theory about how to determine the accuracy of a simulator. This is what we think is important to do in our own work. However, we believe that the approach we outline is applicable for other developers of audible simulation systems and is beneficial for users of these systems.) Authentication is the process of quantitatively determining – by doing scientific, subject-based listening tests – the accuracy of an audible simulation system. The entire purpose of audible simulation is to allow judgments by listening. Authen-

tication, therefore, uses the only valid test of accuracy, and that is to compare listening judgments made on the simulator to listening judgments made in the real world. Listening is at the core of authentication work.

Audible simulation systems are developed to allow sound system designers to interact with simulations of sound systems in the same way they interact with actual systems. In the first part of the paper we will explore this interaction and show how it leads directly to the central question of simulator accuracy. A simulator may allow users to work with the two systems (simulated and actual) in the same way, but there will always be some uncertainty about how close the two systems sound, so in the second part we will explore in detail the consequences of such an uncertainty. In the third part of the paper we will discuss simulator accuracy in detail and propose a strategy for doing authentication. In the fourth part, we will discuss the output of the authentication process. This section will be of interest to potential users of audible simulation systems since it spells out three critical parameters they can ask for from the simulation system's developer. And finally, we will discuss the benefits of knowing the accuracy of an audible simulation system, and we will argue that there are benefits to the audible simulation system developer, the sound system designer, and the sound system customer.

## 1 THE SIMULATED AND THE ACTUAL SOUND SYSTEMS – MIRROR IMAGES

Sound system designers will use audible simulation systems to design sound systems. But, instead of working in real environments they will work in simulated environments that exist only on a computer. In essence, they will replace an actual system by something much more convenient: a computer model. Ideally, we want the way designers (and their clients) interact with the simulated system and judge its sound quality to be the

same as the way they interact with the actual system. In this sense, then, we can think of the simulated and actual sound systems as being ideally like mirror images of each other. In this section, we will explore in more detail what we mean by "mirror images" and introduce the concept of simulation accuracy as a way to tell how similar the two images really are.

## 1.1 Interacting with and Judging the Quality of an Actual Sound System

In order for sound system designers to be successful, they must provide a product – a sound system – with which their clients are satisfied. To reach that goal, designers must know two things: how to judge the sound quality and how to change the sound quality of a sound system.

Designers judge the quality of an actual sound system by listening, and we believe there are a few especially important dimensions of sound quality that affect their judgments. These dimensions are part of the *psychoacoustical domain* in the world of sound systems. Fig. 1 shows this domain and its primary dimensions: tonal balance, loudness, localization, echoes, and speech intelligibility. We say these are important dimensions of sound quality because if some area in the audience does not get approximately the right amounts of high, low and mid tones; or if the system does not play loud enough; or if the sound does not appear to come from the right place; or if there are objectionable echoes, people are likely to complain. And if the system does not provide intelligible speech, people are certain to complain. (We realize that some people may have other dimensions of sound quality that are of special importance to them. For the purpose of our argument here, there is no reason why these dimensions could not be added to our five dimensions.)

Designers know that they can change the sound quality of a sound system by making changes to the signal processing equipment, the loudspeakers, the acoustical environment, and some listener related factors. For example, a change in the amplifier gain results in a change in loudness. A change in sound quality

can be brought about by making modifications to the loudspeaker configuration, such as changing the number of loudspeakers, using speakers with different directional characteristics, or changing their locations or orientations. Sound quality can also be changed by modifying the surface reflection characteristics and even the room geometry. And our location in the room, and our head orientation all affect our judgment of sound quality. In summary, sound system designers change the quality of an actual system by changing one or more physical factors that belong to the *physical domain* in the world of sound systems. Fig. 2 shows what we believe are primary categories in this domain: signal processing equipment, loudspeakers, the acoustical environment, and the listener (related factors).

## 1.2 Designing and Adjusting an Actual Sound System

With these two pieces of knowledge – how to judge the quality and how to change the quality of a sound system – designers adjust physical factors in the actual system to obtain the highest possible sound quality. The process is iterative. First, designers change the settings of one or more physical factors (in the physical domain) and judge the effect on sound quality (in the psychoacoustical domain). They then make new changes and judge the effect those changes have on sound quality. And so on. In this way, designers guide the design towards completion.

The point is that designers use their hearing, together with their experience in adjusting physical controls, in this iterative process. Listening plays a crucial role in the "feedback loop" of this process; without it, designers simply could not do their jobs of guiding sound systems towards their highest possible quality levels.

## 1.3 Designing and Adjusting a Simulated Sound System

Audible simulation systems should allow designers to interact with and judge the quality of simulated sound systems the

same way that they interact with actual systems. What does this really mean? First, designers judge the quality of an actual sound system by judging (at least) the five dimensions of sound quality: tonal balance, loudness, localization, echoes, and speech intelligibility. Therefore, they had better be able to judge the same dimensions on the simulated sound system. And second, they must be able to make changes to the (simulated) signal processing equipment, the (simulated) loudspeakers, the (simulated) acoustical environment, and the (simulated) listener. If designers can change a physical factor in the actual sound system and hear an effect on sound quality, then they had better be able to make the same change to the simulated system.

If designers can do these things (judge and change the sound quality) equally well when using the simulator as on the actual system, then they can also iterate the simulated system design and guide it, too, towards completion. If the two systems really mirror each other, it would make no difference whether designers worked on and listened to the simulated or the actual sound systems. The two systems would be indistinguishable in terms of controls and sound quality, as shown in Fig. 3.

## 1.4 The Concept of Simulation Accuracy

That designers can judge and change the simulated sound system in the same way they can judge and change the actual system is, unfortunately, no guarantee that the two systems perfectly mirror each other. There is no guarantee that the audible effect of changing a certain factor on the simulated sound system is similar to the audible effect of changing the same factor on the actual system. Designers, or any user, need one more important piece of information: they must know about the accuracy of the simulations. How closely does what they hear on the simulator match what they would hear in the real world?

We can look at simulation accuracy as a way of knowing how closely the two systems (simulated and actual) truly mirror each other. If there is a perfect match, simulations would be indistinguishable from the sound of the actual system. Changing a

certain factor in both the simulated and the actual systems would result in the same change in sound quality in the two systems. The systems would mirror each other perfectly in the physical as well as in the psychoacoustical domain (Fig. 3). However, if the two systems do not perfectly mirror each other, a change of a (simulated) physical factor on the simulated system would cause one change in sound quality, while the same physical change to the actual system would result in a *different* change in sound quality. In this case we could say that the systems mirror each other in the physical domain because one can make the same physical changes, but not in the psychoacoustical domain because these changes lead to different judgments of sound quality.

It is this last case that we think should alert potential users of audible simulation systems. If systems do not mirror each other in the psychoacoustical domain, there will be an audible difference between the simulated system and the actual system when it is installed. We believe that it is essential for users to know about any such discrepancy to fully utilize a simulation system. The danger is that if users are not aware of such discrepancies, then they may get unrealistically high (or low) expectations about the actual sound system. Only if users know when the simulated system accurately mirrors the actual system can such expectations be realistically met. And that brings us to the second part of our paper. Here we will discuss the need for a method to determine when we can expect a simulated sound system to accurately mirror the system that we are simulating.

## 2 MOTIVATION FOR DETERMINING THE ACCURACY OF AN AUDIBLE SIMULATION SYSTEM

Simulation systems are developed in an attempt to make the simulated and actual sound systems perfectly mirror each other. However, we know that they may not do that. The only way to know how closely they mirror each other is to determine the

simulator's accuracy. In this section we will focus on why we believe that this should be of primary concern to the audible simulation system developer, the sound system designer, and the sound system customer. We will argue that one of the reasons for determining the accuracy of a simulation system, is because of the severe consequences of not knowing the accuracy.

## 2.1 Simulation Systems Will Occasionally Fail

Meteorologists spend their time making predictions about the weather. They inform us about the predicted temperature, whether they predict it will be sunny or rainy, about storms and so on. Of course, they can not guarantee that the actual weather will match their forecasts. Usually, however, the weather turns out to be pretty much as they promised. But, occasionally, their predictions fail. They predict a sunny day and instead it rains. We (weather customers) use weather predictions because we have, for the most part, learned to trust their predictions. The number of failures in forecasting – that is the number of times we feel that the forecast was poor – is acceptably low. The tool is useful even though it is imperfect.

Like weather forecasting, audible simulations are predictions too. Occasionally the predictions of audible simulations will also fail. When predictions fail, the sound quality of the system actually installed will be different from that of the simulations. Sometimes the actual quality will be better than that of the simulated system. And, sometimes the quality of the actual sound system will be worse than that of the simulated system. The danger in the latter case is that the sound system designer and the client think that they have a high quality sound system (sunny day), when in fact they do not (it rains). Such a situation will inevitably result in disappointed customers. So the question is, how often will users of audible simulation systems accept poor predictions? Will they accept a poor prediction one out of twenty times? One out of five? We do not really know the answer to that question because it is such a new

kind of prediction. But, what we do know, is that having no idea how many failures to expect on a given simulator – whether that be one in a hundred or one in five – is likely to lead to disastrous results. And that leads us directly to simulation accuracy; if we do not know the simulator's accuracy, we will not know when to expect poor (or good) predictions. So, let us address the consequences of such an uncertainty.

## 2.2 The Consequences of Not Knowing the Accuracy of a Simulation System

Let us explore what it means to have a simulation system and not know its accuracy. What are the consequences? If you have no idea how closely the simulations match the sound of the actual system, then, when you audition a simulated system there are three possibilities: the sound quality of the actual sound system may be better than, worse than, or about the same as, the quality of the simulated system.

In the case that the sound quality of the actual system turns out to be better than that of the simulated system, customers are both losers and winners. On one hand, they lose because the system is over-designed, and therefore is not cost effective. On the other hand, they win because they get a better than expected product. In the case that the real system is worse than the simulated system, the consequences are typically more severe because customers do not get the promised high quality product. And, unfortunately, once the sound system is installed there is often no easy fix to improve the sound quality. The only solution may be a new design. So, customers lose because they have a new sound system with poorer-than-expected sound quality and designers lose because their reputations are compromised. And last, if the sound quality is about the same as predicted you could say that the designer blindly gambled on the outcome and won.

Nobody wants to gamble in this business. Designers get paid to ensure, not to gamble. If we proceed to use audible simulation systems before knowing their accuracy, then we are blindly

gambling on the results. We simply will not know how the simulated systems will sound when actually installed. That will turn audible simulation technology into a kind of amusement. It really is fun to listen to the simulations, and it really is fun to explore different design strategies on the simulator. But without this proof of accuracy, we think there exists a real danger that in the end everybody will be cheated of the benefits that this new technology holds. If simulations are not reliable, then designers and clients will fall back on numerical predictions to imagine the sound quality of a proposed sound system. Can we afford to let this happen?

## 2.3 The Argument for Determining the Accuracy of a Simulation System

In essence, we believe there is a chance that we (the audio industry) will seduce ourselves with the marvelous sound effects of audible simulation and forget to question whether these sound effects are related to reality in any useful way. We have to realize that right now there is not one shred of proof that these simulations have any useful connection to reality.

There is an alternative to this reckless course. We can measure the degree to which the simulated and actual sound systems mirror each other. If we do that, users will be able to use audible simulation systems to make judgments they really trust about the sound quality of the system when it is installed. With quantitative information about the simulator's accuracy, they will know by how much the quality of the installed sound system can be expected to differ from that of the simulated system, at least in a statistical sense. With the same information, users can decide for themselves whether they are willing to accept the failure rate of the predictions. Some may even decide to wait until simulator accuracy has improved. After all, if a simulator is not as good as the numbers and graphs we use now, we should stay away from it and use the current methods.

If we decide to go down this path of determining the accuracy of these simulators, we can, with hard work, make audible

simulation a really useful and powerful tool for sound system designers and their customers. We think everyone in the audio community wants to take this path. But, it may not be so obvious just how one would go about actually doing this. And so it is to this subject that we want to turn our attention next.

## 3 THE AUTHENTICATION PROCESS

The previous section argued that we need to determine the accuracy of audible simulation systems. We believe that it is in everyone's interest to determine the degree to which simulated and actual systems mirror each other.

In this section, we will change our perspective and try to look at the situation from the audible simulation system developer's point of view, because we believe it is the developer's responsibility to provide quantitative information about simulation accuracy to users. We will propose a subject-based method – we call it authentication – to quantitatively determine the accuracy of a simulator, and we will describe and explore the implications of that method. Because there are, as you will see, so many experimental conditions involved, and because these experiments include the use of human subjects reporting what they hear, authentication experiments are time consuming and require a multitude of skills. The scope of a program to fully authenticate an audible simulation system is enormous. And yet, these authentication experiments are unavoidable if you want to determine the accuracy of an audible simulation system. Fortunately, it turns out that there is a way to systematically attack the problem, and in so doing get the most important answers first.

### 3.1 Authentication

Let us start by defining the new term for this field – *authentication*. Authentication is the process of quantitatively determining, by doing scientific, subject-based listening tests, the accuracy of an audible simulation system. In authentication

experiments, listeners' responses listening to the simulated sound systems are compared to their responses when they listen to the actual systems.

In an authentication experiment, we determine the simulator accuracy in representing a single dimension of sound quality using a multitude of test conditions. Therefore, the authentication process constitutes several authentication experiments because there are several dimensions of sound quality that we want to be able to judge on the simulator. The strategy we use in performing authentication experiments is to first measure the dimension of sound quality under test using subjects in the real environments. And second, to measure that same dimension of sound quality on the simulation system using the same subjects and the same experimental conditions. Finally, the two sets of numbers are compared. The degree to which they match determines the accuracy of the simulator in simulating that dimension of sound quality under those conditions tested.

Fig. 4 captures the essence of the process of authentication. A subject listens to the actual sound system in its environment. Then the same subject listens to the simulated sound system. The question is: to what extent does the subject make the same judgments in the two situations? (We refer to a single subject here only for clarity. In practice, multiple subjects are always used.)

The essential thing is that authentication uses subjects. The entire purpose of audible simulation is to allow users to make judgments by listening. Therefore, the only valid proof of accuracy is by asking people – in a scientific way – to listen and report what they hear. We have to determine to what degree they hear the same thing on the simulated and actual sound systems by comparing their responses; there is simply no other way to determine if listeners hear the same thing in the two situations.

## 3.2 Scope of Authentication Work

Authentication experiments are required to determine the accuracy of an audible simulation system. However, to completely assess the accuracy would be a lifetime project. First,

there are many dimensions of sound quality that affect people's judgment of a sound system. And second, there are many different physical factors that affect these dimensions of sound quality. Ideally, each one of these factors, and each of the psychoacoustical dimensions they affect, should be tested. Such an experiment is prohibitive. There are simply too many experimental conditions.

In practice, therefore, an authentication experiment is based on a limited number of experimental conditions. With the results of these limited conditions, we would like to extrapolate the results to similar, but untested, conditions. In other words, if the system is accurate in a multitude of typical conditions, then we want to be able to say that it is accurate in all typical conditions. The danger, of course, is that we choose too limited a number of experimental conditions and make an unjustified extrapolation. So let us look at how we can select the dimensions of sound quality to authenticate, and how we can prioritize the physical conditions and test only those that are of primary importance in our work.

### 3.2.1 Selecting Dimensions of Sound Quality

The number of authentication experiments that a developer has to conduct is proportional to the number of dimensions of sound quality that are to be judged on the simulator. In our way of thinking, the psychoacoustical domain for sound system design contains five dimensions: speech intelligibility, tonal balance, loudness, localization, and echoes (Fig. 1). Developers of simulation systems for sound system design must therefore ultimately perform at least five authentication experiments. One to determine the accuracy in terms of speech intelligibility, another experiment to determine the accuracy in the dimension of tonal balance, another in loudness, another in localization, and finally one in echoes. Each dimension of sound quality requires a separate authentication experiment, or possibly experiments.

### 3.2.2 Selecting Physical Factors that Affect Sound Quality

Typically, there are a multitude of physical factors that affect a single dimension of sound quality, and in the ideal situation, all of these physical factors should be included as experimental variables. Such an experiment is probably prohibitive because there are simply too many different factors to test. Instead, experimenters have to choose something more limited. They should select only those factors to vary that are most frequently used to tune actual sound systems. For example, if you were trying to authenticate speech intelligibility, you might choose to vary physical factors related to room geometry and room surface materials since we know that they affect reverberation which in turn affects speech intelligibility. You may also want to include the factors of loudspeaker directivity and speaker layout (centralized versus distributed systems). The idea is that if you have to narrow things down (which you do) then you should choose the conditions you will test carefully, taking the time to ensure that those you choose are ones used most often in tuning actual sound systems.

### 3.2.3 The Challenge in Selecting the Experimental Conditions

The danger in reducing the number of conditions to test in the authentication experiment is that there will be some excluded, yet important, conditions where we will just not know the accuracy of the simulation system. When we choose conditions for the experiment, we by definition exclude others. If excluded factors are used in a particular design, therefore, we have no guarantee that the simulation is accurate for that design. For example, experimenters can choose not to include rooms with a reverberation time of more than four seconds, but if users have to design a sound system in a room with a reverberation time of six seconds, they do not know for sure whether the results from the authentication experiment can be applied in that situation.

Such consequences put a lot of emphasis on the selection process.

If experimenters take too many shortcuts in selecting conditions, the consequence is that there are fewer designs where users will know the accuracy of the simulations. Or, said a little differently, there are more situations where users can not rely on the simulated sound system to be a good mirror image of the actual system. Only if experimenters test a multitude of typical conditions can the results give users the confidence that the simulator can be used safely under conditions not explicitly tested. For example, if a multitude of rooms in the one-to-four second reverberation time range have been tested, users are likely to, and probably are justified in, trusting the simulator for a five second room. In contrast, if the experiment included only one loudspeaker type, users will feel uncertain about using different speakers in their designs. So the more conditions that are used in the authentication experiment, the more confidence users will have in extrapolating the results to conditions beyond those explicitly used in the authentication experiment.

## 3.3 Consequences of Doing Authentication Experiments

Because the scope of (even limited) authentication experiments is huge, it is tempting for developers to try to find other ways of determining the simulator accuracy. However, we strongly believe that there are none. It is simply impossible to determine whether listeners give the same judgment when they listen to the simulated sound system as they will when they listen to the actual system other than by asking them to listen and report on what they hear.

Therefore, authentication experiments have at their heart scientific, subject-based listening tests. These listening tests are time consuming, and require expertise to design and conduct. Authentication experiments measure the response of something very complicated: namely, human subjects. Their judgments inherently have a lot of variability. They respond

differently over time, even when given the same task. And different listeners respond differently given the same task. Consequently, to get precise data the same task must be repeated over and over again. If one takes shortcuts, these experiments are useless simply because the natural variability in the responses of human subjects will obscure any true relationship between the real and simulated conditions.

We have to stress that objective measurements or anything else cannot be substituted for subject-based testing in the authentication process. For example, even though we know that system frequency response has some effect on speech intelligibility, you cannot by measuring the frequency response of the simulated and actual sound systems tell for sure what the difference in intelligibility will be. You have to use subjects. Similarly, you can not measure the impulse response of the actual and simulated sound systems and say that the differences tell you whether the simulator is accurate. You have to use subjects and ask them what they hear listening to both systems.

This is not to say that non-subject-based measurements are never useful. They are very useful as tools for diagnosing problems with a simulation system. For example, comparisons of measured and simulated binaural impulse responses may tell the developer about major discrepancies and defects in the computerized room and sound system model. It is valid and necessary to do such comparisons during simulator development. However, two sets of binaural impulse responses may look different when compared visually. But if they sound the same (in the dimension of sound quality under test), such a difference is not significant in terms of determining simulator accuracy.

In the course of an authentication experiment, subjects give hundreds (sometimes thousands) of judgments. Therefore, when developers have completed an authentication experiment they are left with a huge amount of raw data. Giving users of audible simulation systems all of that raw data is not really informative. One could actually say that it would be more confusing than

useful. So next, we will explore in detail a method for developers to distill the results of authentication experiments, in such a way that it spells out the parameters that are most important to users.

## 4 OUTPUT OF THE AUTHENTICATION PROCESS

One of the most compelling reasons to do authentication is the often severe consequences of not knowing the accuracy of a simulator. However, that developers have done some authentication work is no guarantee that a particular simulator is trustworthy and useful to users. It may be that the simulator is simply unreliable, in the sense that the authentication work shows poor agreement between listeners' responses obtained on the two systems (simulated and actual). Or, it may be that the developer did poor authentication work, so that users are still uncertain about the usefulness of the tool. The point is that when authentication is completed of a particular simulator, potential users of the simulator need to critically examine the output of the authentication process. This leads us to want to say more about the results of authentication experiments, to perhaps give users a better understanding of what they should expect when they are presented with authentication results.

We will therefore change our perspective again and try to explore from the user's point of view a way to evaluate whether a certain simulation system would be beneficial to them. We will argue that by asking for three essential parameters related to simulation accuracy, users will be able to determine whether authentication results apply to their specific design. Moreover, they will be able to estimate by how much the sound quality of simulated systems can be expected to differ from actual systems when installed, at least probabilistically. What developers have to do in order to provide this information is simply to report in a systematic way the design and results of their authentication experiments.

## 4.1 Parameter 1: Dimensions of Sound Quality to Be Judged on the Simulation System

The first parameter a potential user should ask for relates to the capability of the simulator in the psychoacoustical domain. In which dimensions of sound quality has the simulator been authenticated? For sound system design, ideally users should be able to trust the simulator in all five dimensions of sound quality (speech intelligibility, tonal balance, loudness, localization, and echoes). In practice, however, they may have to start with something less, until developers have completed their authentication work. Users will have to make a list with their needs in prioritized order, placing those dimensions of sound quality most important to their work highest on their list. For example, some dimensions may be so essential that a simulator would be useless if it was not authenticated in that dimension. Some users may even have other dimensions of sound quality, such as timbre, that are of special importance to them. When they have completed their list, they then should ask the developer which dimensions have been authenticated, and put those on a second list. They should compare the two lists: their list and the developer's list. If there is a match they can go to the next parameter. If there is a serious mismatch they may decide not to use the simulation system and look for another simulator, or maybe wait for further authentication work to be performed.

## 4.2 Parameter 2: Physical Conditions to Be Used when Designing Simulated Sound Systems

When potential users know what dimensions of sound quality they can judge on the simulator, they then need to know under what range of physical conditions those dimensions can be judged. So, the next parameter relates to the simulator's capability in the physical domain. We are not yet talking about how accurate the simulations are under those conditions (that is the next parameter). We are only addressing the range of physical conditions that were included in the authentication testing. For example, has the effect of changing equalizer settings on tonal

balance been included in the conditions tested? Has the effect of changing reverberation time on speech intelligibility been included? And so on. The strategy we want to use is similar to the one we used when evaluating the capability of the simulator in the psychoacoustical domain, and that is to develop two lists: one with users' needs and one with developers' claims.

Developers can relatively easily complete their list, because it comes directly from the experimental conditions of their authentication experiment. For users to complete their lists, they must consider the range of physical conditions they anticipate using on the simulation system. So each user will, in general, have a unique list. To develop the list, a user must go through the four categories of the physical domain (signal processing equipment, loudspeakers, acoustical environment, and listener related factors), and for each category put down those factors they use most frequently to tune actual sound systems. For example, if you were interested in judging speech intelligibility, factors of primary importance to you related to signal processing equipment could be amplifiers, equalizers, and electronic time delays. For factors related to loudspeakers, you might commonly work with speakers with low directivity, in both centralized and distributed speaker layouts. And so on. The selection process is the same as that used by developers when they select the experimental conditions for the authentication experiment, yet it is different because the perspective has changed; now it is your choice as a user to determine what is important in your work. When your list is complete you compare the two lists: your list and the developer's list to see how closely they match. Because there are so many factors that affect each dimension of sound quality, the two lists will probably never be exactly the same This comparison is crucial and deserves to be discussed in detail.

Based on the comparison of the two lists, you have to determine if you can extrapolate to your work the accuracy of the simulation system obtained in the (limited) authentication testing. This can be a difficult decision to make and the conse-

quences of making a bad decision can result in unexpected (maybe even catastrophic) discrepancies between the sound of the simulated system and the system actually installed. The question we all want to answer is, "When can we make a justified extrapolation?" And, unfortunately, we do not have a complete answer to that question. What we do have is some basic guidelines.

First, there has to be some match between your list and the developer's list. For example, in the authentication work, developers could have included a rectangular 60' x 60' x 30' (18m x 18m x 9m) room with a reverberation time of three seconds. In a specific design you may have a 90' x 90' x 30' (27m x 27m x 9m) rectangular room with a reverberation time of 3.5 seconds. The decision you have to make is whether the results from the authentication experiment (the smaller room) apply to your work (the larger room). In this case you could almost certainly apply the results to your work because the environments are similar in shape and in reverberation time.

Second, the more environments included in the authentication experiment, the more confidence it should give you that an extrapolation is justified. So, going back to our example, the developers could have included in their authentication experiments ten rooms, differing in sizes and shapes and having a variety of reverberation times. If your typical environments fit in among these environments, then you could make a justified extrapolation, even though no authenticated environment perfectly matches your typical environment.

Third, if there is no match at all between your list and the developers' list, then you should be cautious and not make any extrapolation at all. If you do that anyway, you would be gambling with your customer's satisfaction, because you have no guarantee that the simulations are accurate for those systems that you design. For example, if you typically work in rooms with reverberation times around four seconds and the developers only included in their authentication experiment three rooms with reverberation times of about one second, any extrapolation

would be unjustified; the difference is too large in listening experience between two such classes of environments.

These decisions require experience in using audible simulation systems, and because this technology is so new to all of us, nobody possesses this experience yet. Therefore, optimal use of these new systems requires a high degree of cooperation between developers and users until we all have become comfortable determining whether, and how, we will best and most effectively use this technology in our work.

### 4.3 Parameter 3: The Accuracy of the Simulations

By looking at the first two parameters, you should know whether the results from the authentication experiments apply to your work. If they do not apply, you should stop and look for another simulator, or wait for, or even demand from the developer, further authentication work. If they do apply, you can proceed to examine the third parameter: the accuracy of the simulator.

With only the first two parameters, potential users still will not know whether the sound quality of simulated sound systems is similar to the quality of systems actually installed. The way to estimate this is to look at the results of the authentication experiments. So, the third parameter relates to the correspondence in the authentication experiments between listeners' responses listening to simulated systems and to actual systems. This is of primary importance to users, because it will tell them whether they can expect the quality of the two systems (simulated and actual) to be similar, moreover, it will tell them how large or small a difference to expect. We can not categorize simulators simply as accurate or inaccurate. Every simulator will turn out to be accurate to some degree, and it is in quantifying this degree, and discussing its implications, that we are interested. Thus in this section, we will try to focus on what we believe users want to know and how developers can provide this information.

When users make judgments on the simulated sound system they want to know, or predict, what their judgments will be on

the actual system, and what the expected errors are on those predictions. They want to know this error over the entire range of the dimension, for high quality designs, barely acceptable designs, as well as designs with major flaws. It may seem funny, but for a simulator to be really useful it has to be able to sound "bad" correctly. So, to users, two things are important: 1) the level of agreement (found by doing authentication), and 2) the range of the dimension (of sound quality) over which the agreement has been established. Let us illustrate this by using speech intelligibility as an example. Users may want to judge the speech intelligibility of a design by playing standard word lists through the simulated system from which they obtain percentages of correctly understood words. What they want to know is when they play the same word lists through the actual system, what percentages will be obtained. Moreover, speech intelligibility is accepted as being good if the percentages, or scores, are higher than about 90%, fair if they are between about 80% and 90%, and poor if they are below about 80%. So what users also want to know is if they get good, fair, or poor scores on the simulated system, then can they also expect to get good, fair, and poor scores respectively on the actual system.

Let us now look at how developers can provide this information. The results of an authentication experiment should be presented in a way that shows, if possible on one graph, how well responses from the simulated system match those from the actual system. We believe that the most comprehensive way to do this is to plot the results as shown for a hypothetical test in Fig. 5. (We want to stress here the all data in Figs. 5-7 are hypothetical results. No real authentication work has been performed.) The horizontal axis shows the mean values of the scores obtained in the authentication experiment on the simulated systems, and the vertical axis shows the mean scores obtained on the corresponding actual systems, along with the error bars describing the uncertainty of the scores in both directions. This uncertainty arises from the variation in human subject responses on both the simulator and the real system.

The main advantage of presenting the results in this way is that it shows the data from every condition tested and allows the important conclusions to be made by eye. The strength of the relationship between the simulator and reality in the tested dimension is revealed by how tightly the points and error bars cluster about the diagonal line of "perfect agreement". The points themselves are less important than the rough envelope of the outer edges of the error bars. (We will have more to say about these error bars later.) If the simulator tends to give scores too high or too low in any region this will be easy to see as a deviation of several points above or below the ideal line. If the simulator is more accurate in one range than another this will be obvious by a tighter clustering of points along that part of the line. Any extremely poor predictions will stand out from the data, giving a warning that the simulator may be unable to deal well with some physical conditions.

The primary disadvantage of the display is that it does not reduce the experimental results to a single accuracy figure that is easy to remember and compare to other systems. But, in our way of thinking, such a reduction would usually be an oversimplification of the relationship between the simulator and reality. A single number would not tell you whether the error is just scatter, or upward or downward bias (lots of points above or below the line), or a warping of the curve. It would not reveal whether there is more error at one end of the range than the other. To many users, some kind of errors matter more than other kinds. In our view, a side by side comparison, of the kinds of graphs we have proposed, from two different simulators would allow most users to easily pick out the simulator more suited to their needs – even in the absence of a single accuracy figure for each simulator. If it is hard to pick between the two graphs in such a comparison, the simulators are probably about equally good in that dimension of sound quality.

We now return to consider the error bars more carefully. The basic reason for the error bars is that each score of the test is only an estimate of the true score that would be obtained for an

infinite number of trials in that condition. The error bars show a range that the true score is probably within (typically with a confidence of 95%). The more trials we do in a given condition the more we narrow down where the true score is and the smaller the error bars get. The more we know about where the true scores are, the more we can say about whether the true scores from the real system agree with the true scores from the simulated system. This works the other way, too. The fewer trials we do for each point, the bigger the error bars get, and the less we know about where the true scores are, the less we can possibly say about whether the true scores of the real system would agree with the true scores of the simulator. So, we must have the error bars on the graph because their size really does put limits on what we know.

To make this phenomenon clearer, let us consider another hypothetical test of the same simulator we used in Fig. 5. In this new test we have saved time by using only one fourth as many trials to get each score. The results are shown in Fig. 6. Notice that the mean values of the scores are not quite the same and that the error bars have approximately doubled in size, both due to the smaller number of trials. Although the simulator used in the two tests is the same, it seems to be less accurate in the second experiment. Of course the simulator's true, underlying, accuracy has not changed – it is just that the smaller experiment of Fig. 6 tells us less about what the accuracy really is. Naturally, the smaller experiment says less about how well the simulator agrees with reality. In fact, it says so little about the accuracy of the simulator that it would provide most users with little basis for trusting the simulator. This shows why good authentication experiments are so time consuming – developers have to take enough data to shrink the error bars down to a level where the remaining uncertainty about accuracy no longer matters to the users. It is a strength of our proposed graphing format that this effect is so easily seen in the data.

The second important thing to users, other than the degree of accuracy, is the range over which the sound quality dimension

has been tested. Fig. 7 shows hypothetical results from yet another authentication test of the same simulation system, and we can see that the experiment produced no scores lower than 75%. To users, the consequence is that, when they get low simulated scores, they will have no idea what the scores of the actual system will be. Such a simulation system can not warn them about bad designs, unless further authentication is done in the low intelligibility region.

These three examples show that presenting the results of the authentication experiment as shown in Figs. 5-7 is comprehensive because it is fairly easy for users to get all the information they need to judge the accuracy of the simulation system. Users only need to look at a single graph for each dimension of sound quality. These examples have also stressed the importance of including the error bars when presenting results from authentication experiments. Only if the error bars are small can we possibly conclude that a simulator mirrors the sound quality of real systems well in any particular dimension of sound quality. Without the error bars, we can tell nothing at all.

If users are satisfied with this third parameter, that is, they are satisfied with the accuracy of the simulation system, they will have, when they combine this third parameter with the first two, a good understanding of the strengths and weaknesses of the simulation system: they will know the dimensions of sound quality that can be judged, they will know the physical conditions where those dimensions can be judged, and they will know how accurate the simulations are under those conditions. These three parameters give a thorough description of the simulation system and provide a framework that allows users to evaluate the usefulness of the simulation system for their work. If they are satisfied with all three parameters, the simulation system will be useful to them and simulated systems will behave acceptably close to actual systems when installed.

## 5 THE BENEFITS OF KNOWING THE ACCURACY OF A SIMULATION SYSTEM

The previous sections have argued, that before we start using an audible simulation system for design or communication purposes we need to trust the tool. Developers need to trust it, so they do labor extensive authentication work, to determine to what degree simulations represent the sound of what they are simulating. Users need to trust it, so they critically examine the output of the authentication work, to determine if the simulator at all will be useful in their work. The point is that nobody would be involved in this process, if they do not envision the profound benefits this technology holds, once it has been proven trustworthy. So, in this section we will focus on the benefits of having complete authentication results, results which satisfy users in all three parameters in the dimensions they care about.

There are many benefits to us, as developers and researchers. First, as developers, we have the fun and pleasure of taking such a technology from a theoretical idea to a practical design and listening tool. But, there are also benefits to us, as researchers. First, we will learn the strengths and weaknesses of our computer modeling tools. One of the problems in modifying existing tools is that it is so difficult to evaluate whether our "improvements" have the desired effect, simply because we have to evaluate their effects by looking at numerical results. With authenticated audible simulation technology, that will no longer be a limitation, because we will be able to judge the effects by listening tests. That alone will be a fundamental aid in developing future improvements of computer modeling tools. Second, there are many new kinds of psychoacoustical experiments we will be able to conduct once some authentication work has been completed. For example, we will be able to study the audible effects of changing the early reflection pattern, without having to make any physical changes in order to create the different patterns. Audible simulation provides a whole new set of possible research directions.

There are numerous advantages for sound system designers when they know the strengths and weaknesses of their simulators. First, they will be able to design sound systems on the computer and judge sound quality by listening, the same way they proceed when working with an actual system. No component used in sound system design is perfect, but designers can create very high sound quality if they know the limits of their components. Similarly, an imperfect simulator can contribute to a better design, but only if the designer knows the limits of the simulator. If they know the accuracy of their simulators, they will be able to design better sound systems because they can listen to their progress as they design. And second, simulators will provide them with a high degree of confidence in their designs, because they will no longer have to make conclusions based on numerical predictions alone. For example, determining the annoyance of a late-arriving strong reflection by looking at an impulse response is currently very difficult; one always wonders while staring at the amplitude and time of arrival, and its relationship to other features in the impulse response, to what degree it might be annoying. With a good audible simulation system, designers can simply listen and decide for themselves if it is annoying. And so can their clients. This confidence will be an enormous advantage when designers go to their customers to convince them of the value of a quality design.

There are also benefits to clients. Currently, even though sound system designers do their best, they are forced to give their clients the difficult job of evaluating the proposed sound system by looking at numerical predictions. Graphs and figures of acoustical terms like sound pressure levels and sound coverage, may mean a lot to us, but to expect customers to understand them on anything but a relatively shallow level is naive. In the future, with good audible simulation systems, clients will be able to listen to the design and that will make their decision about its quality much easier. A customer will become a full partner with the designer in assessing the quality of a proposed system, and that is the way it should be.

## 6 CONCLUSION

In the last few years, a number of research and development efforts have been devoted to audible simulation technology. The potential benefits of being able to listen to a computerized model of a sound system in a room are clear and profound. However, we believe there exists a real danger that we (the audio industry) will forget to ask ourselves seriously whether the simulations bear any resemblance to the sound of the environment they are simulating.

We have in this paper focused on what we believe is a crucial part of audible simulation: quantitatively determining the accuracy of simulation systems before we use them for design or communication purposes. If we do not take the time to do this, we will lose control over this promising technology. Determining the accuracy of a simulator is the only way to know if the quality of simulated sound systems will have anything to do with the quality of the systems when they are installed. Without this important information we can simply not make any intelligent judgments by ear about how the system will sound when installed.

We believe that it is possible and practical to quantitatively determine the accuracy of a simulation system. We have proposed a strategy – we call it authentication – to determine by doing scientific, subject-based listening tests, to what extent people hear the same thing in the simulated environment as they hear in the real environment. The entire purpose of audible simulation is to allow listeners to make judgments by listening, so the only valid proof that it works is by comparing listeners' judgments listening to simulated and actual systems. The essence of authentication is listening.

When developers have completed their authentication experiments, they can structure the results in such a way that they describe three parameters related to simulation accuracy; these are essential for allowing users to determine whether a given simulation system can be useful in their work. The first param-

eter relates to those dimensions of sound quality that can be judged on the simulator, the second to the physical conditions under which sound quality can be judged, and the third to the accuracy of the judgments. Users each have their own unique requirements, so what they must do is to characterize their needs – in terms of these three parameters – and see if they match what is offered by a given simulation system. In this way users can ensure that they do not get a simulation system that turns out to be a useless, or worse misleading, tool.

When we have done authentication, we can start using audible simulation systems wisely and carefully to serve our customers. If we do no authentication, we can still dazzle them and ourselves with impressive sound effects. At least for a while. Which course do you think our customers want us to take?

## 7 REFERENCES


[1] J. Martin, "A Binaural Artificial Reverberation Process," presented at the 91th AES Convention (New York, 1991), Preprint #3121.

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

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

[4] A. Mochimaru, "Evaluating the importance of source directivity, frequency, and phase response in the auralization of sound system designs," presented at the 124th ASA Meeting (New Orleans, 1992), Paper 3aAA4.

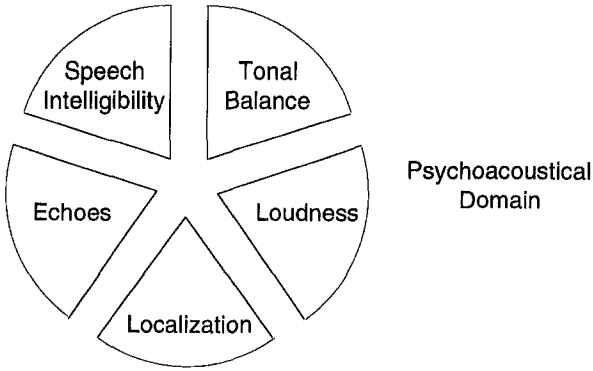


Fig. 1. In sound system design there are a few important dimensions of sound quality that affect people's judgment. These dimensions are parts of the psychoacoustical domain. The figure shows what we believe are the primary dimensions of this domain: speech intelligibility, tonal balance, loudness, localization, and echoes.

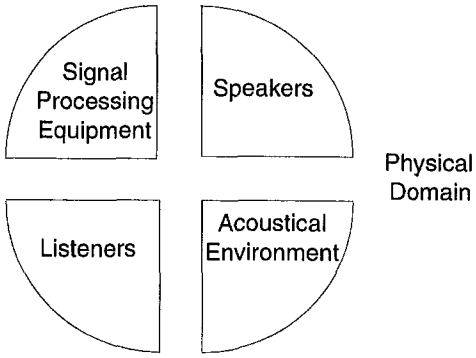


Fig. 2. When we adjust actual sound systems, we make changes to certain physical factors that we know have an audible effect on sound quality. These factors are parts of the physical domain. The figure shows what we believe are the primary categories of this domain: signal processing equipment, loudspeakers, the acoustical environment, and the listener. Within these categories are the factors that we physically adjust. For example, in the signal processing equipment category, we might adjust amplifier gain, equalization, and electronic time delays.

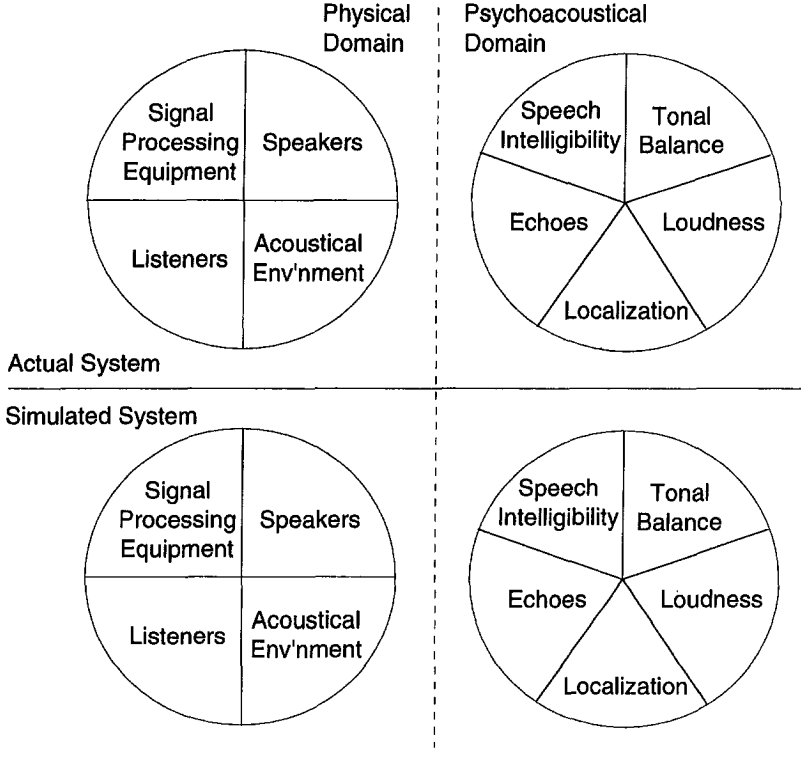


Fig. 3. Audible simulation systems are developed to allow us to interact with and judge the quality of simulated sound systems in the same way that we interact with and judge the quality of real sound systems. In this sense, the simulated system is ideally a mirror image of the actual system. In order to be useful, the simulator must mirror the real system in both the physical and in the psychoacoustical domains.

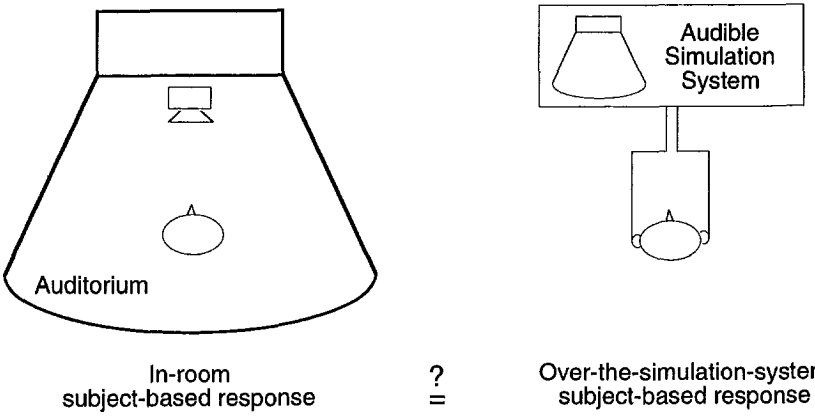


Fig. 4. This figure captures the essence of authentication, the process of quantitatively determining, by doing scientific, subject-based listening tests, the accuracy of a simulation system. First, a subject listens to the actual sound system in its environment (left) and then the same subject listens to a simulation of that system (right). What we want to know is to what extent the subject hears the same thing in the two situations.

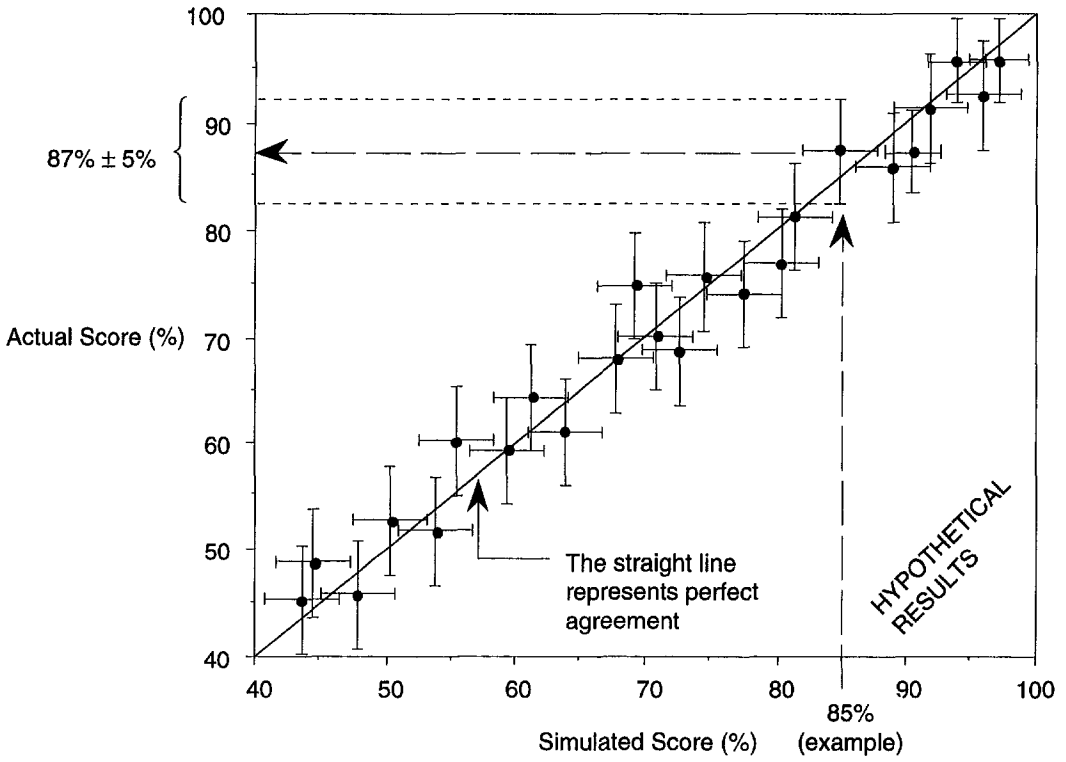


Fig. 5. The figure shows the results of a hypothetical authentication experiment. The horizontal axis shows the mean values of the scores obtained in the authentication experiment on the simulated systems, and the vertical axis shows scores obtained from the corresponding actual systems, along with error bars describing the uncertainty on the scores in both directions. From this figure users can estimate how much actual scores can be expected to differ from simulated scores. For example, a simulated score of 85% appears to correspond to an actual score of about 87% ± 5%, that is with high confidence (typically 95%) the actual score will be in the interval from 82 to 92%. Users can also easily tell, by examining how closely points fall near the diagonal line (which represents perfect agreement between the simulator and actual systems) how accurate the simulator is in various ranges. Furthermore, by looking at the error bars, users can evaluate how much the experiment actually revealed about the true accuracy of the simulator. Only if the error bars are narrow, can we possibly conclude that a simulator mirrors the sound quality of real systems well. If the error bars are wide, we do not really know, and without the error bars, we can tell nothing at all.

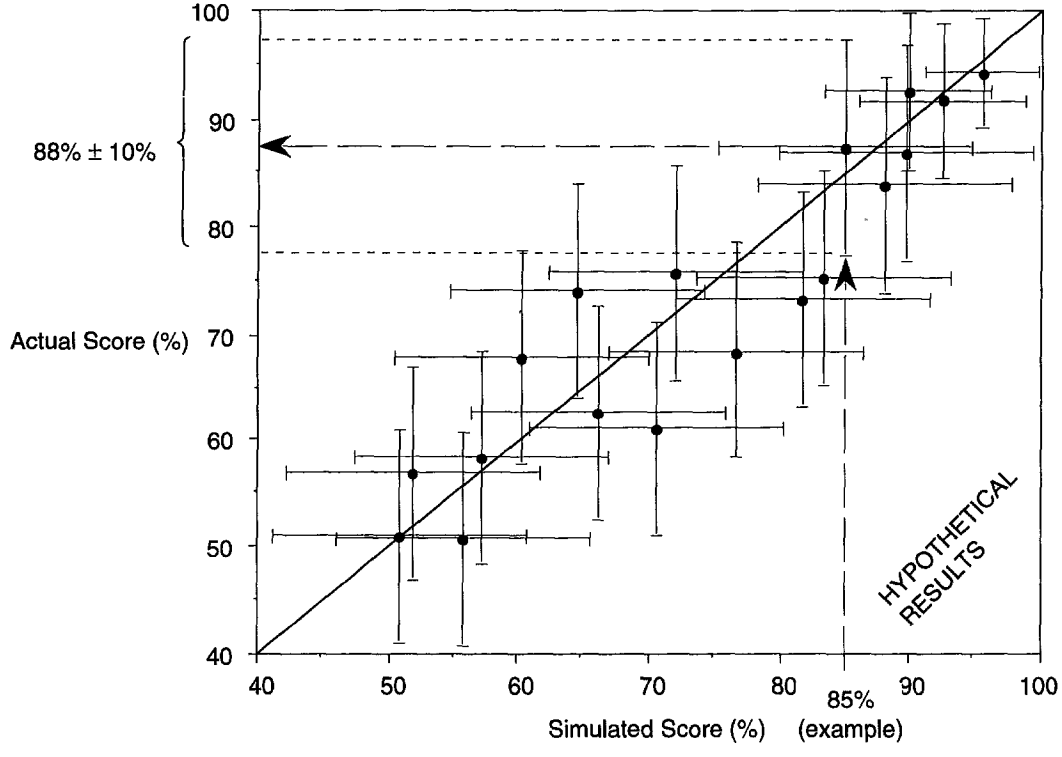


Fig. 6. This figure belongs to the same family as Fig. 5, and shows results from another hypothetical authentication experiment in speech intelligibility. However, fewer trials were used here than for the testing presented in Fig. 5, and the error bars are, therefore, wider. A simulated score of 85% appears to correspond to an actual score of about 88% ± 10%, that is with high confidence the actual score will be in the interval from 78 to 98%. In this case, we are less certain about the behavior of the actual system than we were using the data in Fig. 5, simply because the number of trials in the authentication test was smaller.

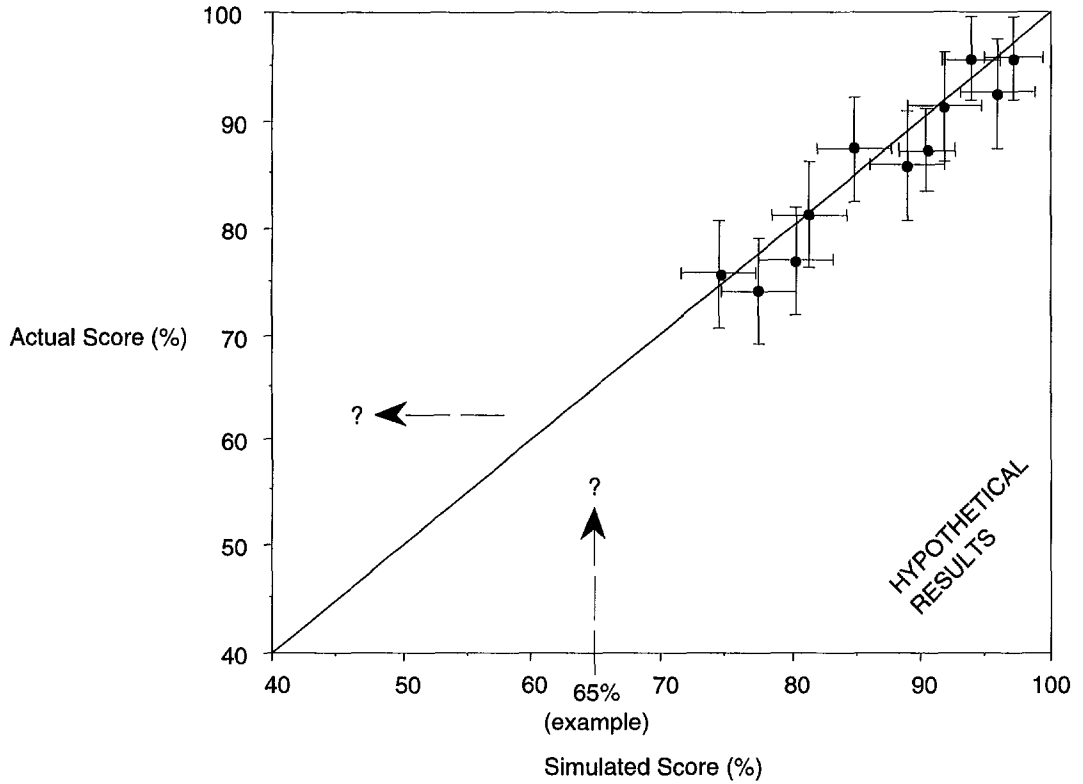


Fig. 7. This figure also belongs to the same family as Figs. 5 and 6. The results are from another hypothetical authentication experiment in speech intelligibility. Notice that the experimenter has no data establishing the simulator accuracy for scores below 75%. If users obtain simulated scores in that range during a design, they would not know anything about the scores of the actual system. This simulator may not be able to tell designers if they really have a bad design.