AUDITORY AND NON-AUDITORY FACTORS THAT POTENTIALLY INFLUENCE VIRTUAL ACOUSTIC IMAGERY

DURAND R. BEGAULT

San José State University Foundation/Human Factors Research and Technology Division
NASA Ames Research Center, Moffett Field, California
db@eos.arc.nasa.gov

Both auditory and non-auditory factors affect the ability for a sound designer to manipulate auditory localization, distance, and environmental context perception. The influence and possible detrimental effects of room acoustics, listening position, and spatial and temporal asynchronies are reviewed. Different approaches to spatial evaluation are reviewed in light of the demands of the application context.

INTRODUCTION

Research into the localization of virtual acoustic stimuli is often done in isolation from both acoustical and non-acoustical factors that are present in real-world environmental contexts. The fact that the commercial industry often separates “audio” from “visual” engineering is a reflection not only of hardware expertise but also of the specialization of psychophysical knowledge. For example, our collective knowledge of psychoacoustics far exceeds our knowledge of the multi-modal interaction between sight, audition, and tactile sensations. However, technology developments benefit considerably when it is possible to predict how perception of an event within one sensory modality is affected by simultaneous presence of stimuli from other modalities.

The study of multi-modal interaction, primarily between audition and vision, has received increased attention with the development of virtual reality systems, home theater, gaming, and teleconferencing. Several authors have reviewed inter-sensory interactions from a host of different perspectives [1-5]. This paper reviews selected aspects of both visual and infrasonic stimuli interaction with sound and their potential influence on virtual acoustic imagery. Specifically, the focus is on the perception of spatial attributes under both headphone and multichannel audio-visual playback. The interaction of cues from other sensory modalities are proposed to act as potential sources of “noise” that are detrimental to both localization and subjective evaluation of an acoustic space. In that sense, it is necessary in the near future to question the emphases used in subjective evaluations for the development of virtual acoustic and audio-visual simulations.

First, a review of localization performance of 3-D sound under headphone conditions is reviewed. Headphone playback is considered optimal for reproducing 3-D spatial acoustic imagery because of their relative immunity to acoustical detriments and the consequent power with which a sound recording engineer can predict resulting perceived spatial attributes. The related studies tend to be centered on specific aspects of HRTF manipulation and the subsequent effect on accuracy of localization or other performance measures. Simulations can be further improved with the addition of reverberation and head-tracking cues. Second, a brief review of subjective localization studies involving audio-visual interaction is presented. These studies contrast with localization performance studies in that the spatial qualities assessed are quite different. The application focus of the audio-visual interaction analysis is on research pertinent to multichannel home theater systems. Finally, some of the potential interactions between audio, visual and vibro-acoustic sensation are addressed as they influence the perceived quality of spatial simulation. This includes an assessment of the interaction between the listening-viewing room and the simulated space.

1. LOCALIZATION STUDIES OF 3-D SOUND FOR HEADPHONES

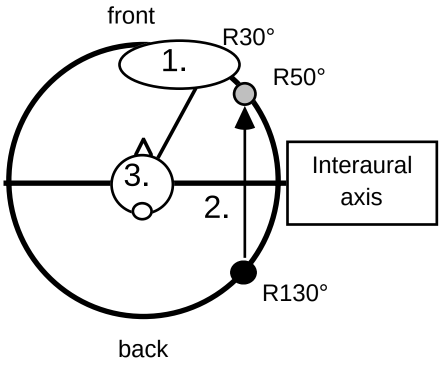
Headphone-based displays allow the greatest degree of control over the location of a spatial source, and are essential to applications where performance of a subject in a specific task is involved. Additionally, the influence of background noises in the listening room can be eliminated. In these regards, headphone playback is considered an “optimal” condition for the reproduction of 3- D sound. Although some of the signal processing techniques for producing cross-talk cancellation can allow loudspeakers to deliver some spatial effects that are not possible with normal intensity stereo, headphones remain the playback medium of choice for maximal control of spatial imagery.

Auditory localization studies have for the most part investigated localization accuracy and the degradation in localization performance (or lack thereof) as a function of one or more independent variables. This objective measure of human localization performance is quite different from an inquiry into a subjective evaluation of quality or “realism” of the spatial aspects of a sound. A similar dichotomy between performance and “immersion” was discussed previously concerning virtual reality [6, 7]. For example, in a virtual reality application, a subject’s performance in manipulating an external virtual object with the hand may be of greater importance than the subjective realism of the object itself, or how “present” the subject feels they are in a synthetic environment.

The reproduction of 3-D sound over headphones can cause one or more of the perceptual errors outlined in Figure 1. The term ‘error” refers to the mismatch between the intentions of the system designer and the resulting percept of the listener. Usually, for headphone displays, the concern is with localization blur, reversals, and problems with externalizing the stimuli.

Many investigations have focused on performance as a function varying some parameter of the head-related transfer function (HRTF) itself. A significant problem for the implementation of 3-D sound systems is the fact that spectral features of HRTFs differ between individuals, and therefore localization errors increases when listening with what are termed “non-individualized HRTFs” [8-11]. A great deal of research has also been devoted to “modeled” or “data-reduced” HRTFs that yield equivalent performance to personal HRTFs in localization tasks (see, e.g., [12, 13]). The mitigation of reversals and unexternalized stimuli emphasizes spatial transformations caused on high-frequency components of incoming sound sources (> 5 kHz), although it has been well established that low-frequency information dominates localization [14]. This is particularly important with regards to loudspeaker-based “cross-talk cancellation” applications of 3-D sound because the time relationship between arriving signals can be altered quite significantly with small movements of the head.

Figure 1: Three examples of localization error in



headphone studies. 1.: a target at 30 degrees with increased width (“blur”) and biased to the left 2.: a back-front reversal of a 130 degree target heard at 50 degrees. 3: an unexternalized target heard at the edge of the head

The performance studies of localization also have been related to application-specific tasks, such as visual search and speech intelligibility using the “cocktail party effect” advantage. Direct application of visual search studies are for highly specialized areas such as aviation safety; detection accuracy and speed as the usual dependent variables [15-17]. “Cocktail party” applications are driven by measures to show improvement in speech intelligibility while listening to multiple channels, and are probably less dependent on perceived location than the net interaural decorrelation between multiple maskers and signal [18-22].

Two other factors are frequently cited as means for improving localization within a 3-D audio headphone-based display: simulation of reverberation cues, and head motion cues. A NASA-sponsored study is in progress by the author where these effects are studied in a direct comparison of their efficacy in improving localization of speech stimuli. (These data were not ready at time of the present publication but are presented at the conference and in an upcoming paper).

Several studies have shown that head movement cues can improve localization ability and reduce the number of reversals [23-25]. Listeners apparently integrate some combination of the changes in ITD, IID, and movement of spectral notches and peaks that occur with head movement over time, and subsequently use this information to disambiguate, for instance, front imagery from rear imagery. Reverberation has been shown to dramatically increase the externalization of stimuli relative to non-reverberated stimuli, in one case, from 2% to 90% [26, 27]. It may possible to mitigate reversal errors by establishing a "cognitive map" of the acoustical features of a reverberant space cues

Training subjects by having them adapt to listening to non-individualized HRTFs over an extended period has been suggested to improve localization of virtual acoustic imagery [28, 29]. There has also been exploration of the possibility of synthesizing "supernormal" localization cues with larger ranges of interaural difference than normal cues, thereby improving the ability to resolve spatial locations [29, 30]. Although these techniques may allow localization improvement, it is important to recognize that it is possible to exceed the capacity of the listener's localization ability with arbitrary expectations. Auditory localization under everyday conditions is relatively poor compared to vision, with the error in large-scale studies cited by Blauert from 4-10 degrees in the horizontal plane and even higher for elevation [31]. The fact that the percentage of localization "reversal" errors has been cited to be around 8% under normal listening conditions and as high as 40% under 3-D simulation conditions is suggestive [10, 32, 33].

The most accurate 3-D sound localization seems to require active, attentive listening, in the absence of distractions from undesired visual, auditory and tactile sources. The influence of cognitive cues, memory, and associations must also be a controlled factor. Accuracy of simulation also requires veridical head movement cues and realistic simulation of the environmental context [34, 35]. Determination of salient acoustical parameters for rooms and other types of environments will require a great deal of additional research on the perceptibility of various acoustical attributes that make up the physical nature of these spaces. Although there are many studies of early reflection thresholds [36-39], results from investigations of directional perception of early reflections seems to suggest only a very general sensitivity under many conditions (see, e.g., [40]).

2. SUBJECTIVE EVALUATIONS OF 3-D SOUND

The types of localization performance measures described above are driven by the optimal conditions offered by headphone reproduction, where greater control is possible over interaural differences and spectral cues. These are most relevant to applications in specialized contexts such as virtual reality, aviation, and demanding human-machine interfaces. By contrast, the context for multi-loudspeaker audio-visual applications such as home theater and teleconferencing present a number of variables that can detrimentally influence localization performance. Speaker cross talk, modal interactions at low frequencies and listening position differences make prediction of a specific azimuth and elevation nearly impossible, although it is still possible to communicate less specific aspects of the intended spatial experience. Consequently, perceptual evaluations must place greater emphasis on the subjective realism or "quality" of spatial percepts. Localization estimate studies are considered to be in a completely separate domain from investigations of spatial attributes of audio quality in [41]; this reference provides a good overview of the current state of subjective responses to spatial stimuli.

In localization tests of 3-D sound, estimates of virtual sound source position are highly reified. A perceived position is indicated in terms of its azimuth and elevation (and sometimes distance) by verbal estimation, by pointing, or other methods. The estimates tend to be made from an "egocentric" perspective, where the position is evaluated relative to the perceived location of the listener. By contrast, evaluations of localization in audio-visual interaction contexts must use "object-oriented" (exocentric) judgements, where positions produced by two modalities are estimated relative to one another. The elicited response focuses on the overall "quality" of the perceived spatial relationship between images, not on perceived location per se. For instance, Hollier and Rimell summarized a model of multi-modal perception as depending on, among other factors, the degree of "quality-mismatch" between audio and visual components [3]. This approach is in line with the general trend for design manufacturers to include sound quality assessments in the design of a variety of consumer products.

While a great deal of work has gone towards establishing which spatial criteria and types of audio-visual interaction are most pertinent, it remains difficult to establish consistent criteria between subjects, rooms, program material, or reproduction systems. Specifications for standardized listening-viewing test procedures are essential to reign in experimental variability. Subjective evaluation of the spatial reproduction of auditory-visual media are addressed in several International Telecommunication Union (ITU) recommendations [42]. Subjective listening-viewing tests are recommended to use five grades for quality assessment (1-5, from “bad” to “excellent”); five grades for evaluating attribute impairment (1-5, from “very annoying” to “imperceptible”); and seven grades for comparative studies (-3 to 3, with –3 as “much worse” and 3 as “much better”). These techniques are outgrowths of listening test techniques that have developed over the years in an attempt to quantify the multidimensional aspects of overall audio system quality or of low bit-rate codecs [43, 44]. The caveat of course is that the more the real world listening situation departs from this standard, the greater the possibility will be for mismatch between what is actually experienced and the findings of a listening test.

ITU recommendation BS.1284, “Methods for the subjective assessment of sound quality- general requirements” contains a category for evaluation of transmission artefacts involving “distortion of spatial image quality” [42]. This is explained as involving “all aspects including spreading, movement, localization stability, balance, localization accuracy, changes of spaciousness.” Since all attributes of spatial listening are pertinent, there is a need for further specification in other recommendations.

Related recommendation BS.1116-1 gives guidelines that are more specific for assessments of spatial imagery for both two-channel stereophonic systems and multichannel systems [45]. For two-channel stereophonic systems, stereophonic image quality is related to “differences between reference and object in terms of sound image locations and sensations of depth and reality of the audio event.” For multichannel systems, the subjective criteria are categorized in terms of “front image quality” (the localization of frontal imagery, image “quality” and “losses of definition”) and “impression of surround quality (spatial impression, ambience or special directional effects).

Of particular interest are the differences in spatial criteria between these systems, although the same objective spatial image may be reproduced. Although stereo is capable of several of the qualities given to multichannel sound, systems are evaluated in terms of the way they perform best. Because one is surrounded from both front and rear by loudspeakers, it becomes possible to easily manipulate spatial impression or ambience in terms of their relationship to perceived auditory width and envelopment.

The related ITU recommendation for multi-modal interaction is BS.1286, “Methods for the subjective assessment of audio systems with accompanying picture” [46]. Here, the attributes for multichannel sound include the front image quality and impression of surround quality attributes mentioned above, as well as the “correlation between sound and picture images. These include “correlation of source position derived from visual and audible cues (including azimuth, elevation and depth)”, and “correlation of spatial impressions between sound and picture.” In practice, the definition of correlation tends to focus on overall impressions of spatial quality “matching.”

Correlation is difficult to define for certain program material where the visual and audible cues are without a common reference. An example is the soloist performing on a grand piano in a concert hall. The audio spatial perspective provided on modern recording techniques of the piano refers to neither the sound heard by the pianist, by the audience, or by someone sitting on stage in 3 ft in front of the lid or with their head 3 in over the sounding board. It is rather a synthetic blend of direct and diffuse sound that yields aesthetically satisfying stereo imaging, where different pitches correspond roughly to a variance in azimuthal position and image width that are suggestive but no means representative of any listener’s perspective. Now, the accompanying visual image, if it is fixed at one location, will certainly not correlate in any literal sense to the recorded sound. The situation becomes more complex when the pianist is viewed from multiple distances and angles in various cuts. For musical and intelligibility reasons, the level remains constant independent of different virtual viewing distances.

In many multichannel audio and audio-visual studies, attributes related to absolute spatial image location are not emphasized, due to the difficulty of precision sound localization reproduction with multiple speakers. Instead, the attributes focus on acoustical parameters associated with the characteristics of diffuse sound, especially those associated with concert hall research [47]. The quality of ambient sound reproduction is often investigated in terms of subjective envelopment (sensation of being surrounded by a source) and auditory spatial width (ASW- the perceived extent of the sound source- see, e.g., [48]). Each of these attributes is a natural consequence of a multichannel reproduction environment in a small-to-medium room, and allows some independence in seating position. Envelopment is a function of the totality of speakers (i.e., those source locations that surround the listener), and depends on the location and number of low-frequency drivers and their interaction with the room[49]. ASW is a function partly of the extent of frontal loudspeakers or other factors that contribute to the relative decorrelation of the audio signal arriving from the front.

It has been pointed out that there is a seeming contradiction between concert hall acoustic studies, where ASW is found to be desirable, and localization studies, where the reflections that cause ASW make localization less precise [41]. A subjective evaluation of spatial qualities would presumably be highest when there is a ‘realistic” or “natural” ASW associated with a particular environment and a sound source location, and not the ASW heard in an anechoic chamber. In other words, it is necessary to evaluate two aspects of spatial perception simultaneously. The technique of combining several spatial attributes for a subjective response along a single dimension is practiced in several studies. For example, in one multichannel sound study subjects evaluated “spatial sound impression” in terms of envelopment, and depth and width of frontal images [48]. Another study had subjects rate “image width, position and motion” for audio-visual coordination and “the spatial naturalness of the whole system, and to consider how true to life the projected presentation is in terms of spatial reduction” in terms of their effect on perceived “naturalness” of the presentation [50].

In discussing the success of 3-D sound rendering, the author has used a simple communication chain of source-medium-receiver in order to describe the differences between the rendered spatial imagery in the recording studio or sound stage, and that heard by the listener [51]. Any differences between source and receiver are explained as a source of undesired noise. However, while this perspective applies to critical human-machine interfaces or applications where performance is an issue, it ignores what could be called the “participatory” aspect of consumer audio. Commercial recording engineers must frequently adjust to changes in spatial imagery between different control rooms and loudspeakers. The small reference monitors popularly used with mixing consoles provide a relatively consistent but by no means preferable presentation of spatial attributes.

The most effective way of minimizing mismatches between source and receiver would be to match the acoustics and loudspeakers between the listening room to the control room. However, this is seldom practiced, in spite of the fact that this would be the most faithful reproduction of “reality.” In effect, the consumer is expected to participate in the final formation of spatial imagery from not only perceptually, but also by creating a design and context for sound playback. This is evidenced by the difference between configurations of speakers, listening position, and other individual differences found in homes, automobiles, etc.. Although it is possible to adjust the timbral aspects of home theatre sound to a standard “house curve” as proposed in [52], it seems inevitable that spatial attributes will change significantly. Consequently, subjective evaluations of multiple attributes of spatial quality are inevitable.

The success of cinema and would have been greatly effected if sounds were required to be co-located spatially with their images. In one review of sound space in cinema, several examples are given of how, prior to 1930, there was a special concern amongst influential technicians to maintaining a “natural” proportion between image and sound [53]. Multiple channel speaker systems were proposed such that the sound could be switched to as close to the position of the visual image as possible. In 1915, Amet patented a “method of and means for localizing sound reproduction,” a speaker switching system for theatrical presentation whose purpose was so that “.. the audible actions may be localized to correspond with localized visual actions, so that if both actions are simultaneously are produced they may truly represent the original production” [54].

A concern for a “naturalistic” correlation between vision and audition can thus be seen from the beginnings of the industry. This extended to distance perception as well In 1928, a film technician writing in *American Cinematographer* stated that viewers would not accept a lack of auditory perspective because their eye/ear coordination would not allow them to [53]. Joseph Maxfield authored several articles in the 1930s that argued for a single microphone placed near the camera’s line of sight for properly capturing distance cues. He offers in one article “…a study of methods of controlling some of the factors available to the engineer in sound recording and photography in such a manner that a pleasing illusion of reality is created”, and emphasizes that his techniques are for the purpose of insuring that the sound appears coming from the visible source on the screen [55].

The eventual development of new technologies such as lightweight, boom mounted microphones allowed the cinematic experience of distance to be subjugated to the needs of speech intelligibility, eventually yielding the ‘god’s ear’ perspective that keeps most dialogue in the foreground, independent of visual location. Modern multichannel sound is then largely to create an artificial space, and the participation of the listener mentioned earlier further causes the search for a subjective response based on “nature” difficult. Overall, spatial correlation between audio and visual in home theater is worth evaluating, but more to have a comparative reference between systems than an absolute measure of localization akin to the headphone studies discussed previously. It may also allow a historical reference to multimedia systems of the future; differences in our subjective preferences may evolve as we learn from new forms of multimedia experience.

3. POTENTIAL MITIGTING FACTORS FOR SPATIAL PERCEPTION IN AUDIO-VISUAL MEDIA

Several mitigating factors can affect both the precision and the quality of spatial sound reproduction as well as audio-visual interaction. To begin with, the context in which audio and audio-visual playback of loudspeaker sound is experienced must be addressed. Next, the effects of spatial and temporal asynchrony are discussed.

3.1 Background acoustical and visual noise

In real world conditions, there is an acoustical coupling between the listening and viewing environment and the simulation environment. The visual simulation environment is constrained by the physical size of the image and the degree to which the visual field of the viewer focused on it to the exclusion of other visual imagery from the real environment. Factors such as lighting and architecture can be very influential on a perceptual environment. Even the visual identification of the type and manufacturer of a loudspeaker can bias overall quality judgements, and so would likely influence judgements of the success of a spatial rendering [56]. It is very difficult to separate the virtual world from the real world in which it is experienced, in spite of the best attempts.

Many of the psychophysical signal detection-type studies examining the influence of visual stimulation on auditory detection and vice versa are reviewed in [1]. In general, listeners are able to more easily detect an auditory signal with the presence of a visual stimulus, while the sensitivity to a visual stimulus may be increased or decreased by the presence of acoustic stimuli. The overall conclusion to be made is that the sensitivity to audio-visual events is increased under multi-modal conditions.

One of the main factors that can influence speech intelligibility and perceived spatial quality of audio is the presence of auditory “maskers,” that is, background noise or other noise sources that are not part of the program material. Background noise levels are seldom the same in everyday listening environments when compared to each other or compared to the control room where the imagery was produced. A convenient way to characterize background noise levels in various environments is indicated by the noise criteria curves (NC) or one of its several variants [57]. NC curves are frequency dependent contours used to establish the relationship between background noise and the target criteria for various types of rooms. The curves reflect the equal loudness contours in that a particular curve allows more energy in lower frequency bands than in higher frequency bands.

A typical listening room in a home without any internally generated noise from children, dishwashers, or noisy air handling devices might be equivalent to an NC 35 level, while a recording studio might be built for NC 10 or better. The ITU-R BS.1116-1 recommendation indicates that background noise levels for subjective evaluation be no more than ISO NR 15 (ISO NR 15 is a slightly more stringent curve than the curve for NC 15). These are far quieter than the background noise levels of concert halls or more importantly of private residences. Obviously, the presence of everyday noises in home environments serves to mask many spectral cues as well as reverberation and ambience cues as a function of overall playback level.

3.2 Acoustical factors of the listening-viewing room

The potential for the detrimental interaction of room acoustics is well known (see, e.g., [58]). Specifically, low-frequency modes and undamped early reflections can affect the frequency and spatial imaging produced by loudspeakers. Different configurations of the sound source (loudspeaker position and directivity), listener, and their context (the listening-viewing room) can result in different relative balances at the receiving position of both direct sound and early reflections. It can therefore be reasonably assumed that these detrimental effects would carry over into subjective evaluations of the quality of audio-visual media.

The nature of these interactions has been dealt with recently in a home theater context. The experiment explored effects of loudspeaker directivity and listening position on several spatial attributes in a “surround experiment” and a “frontal experiment” [50]. The surround experiment had subjects evaluate envelopment and directional detail for rearward sound events, and the spatial naturalness of the system compared to real-life experience. The frontal experiment had subjects evaluate how coordinated were sound and picture *collectively* in terms of image width, position and motion; how correct/how much was the sensation for acoustic space; and how “natural” the projected presentation compared to real-life experience. The common conclusion for both experiments was that loudspeakers with relatively higher directivity were superior for spatial reproduction and for picture-image correlation, because they provided less excitation of the room acoustics. This supports Holman's measurements of early reflections from standard versus restricted directivity loudspeakers; the restricted directivity loudspeakers were more immune from early reflections [52]. Overall, it must be concluded that room interactions and potential detriments are at least important for audio-visual interaction as for purely audio presentation.

The effects of room modes becomes apparent at low frequencies especially in small rooms, where the number of modes is low and are highly separated, thereby causing differential interaction with loudspeakers and seating position. These are heard as resonances or tone colorations, and sometimes as a low-frequency "boominess." The "Schroeder frequency" indicates the cut-off frequency at which a room cannot be adequately described by reverberation time but instead by its "modal response:.

$$Fc = K\sqrt{\frac{RT}{V}} \qquad (1)$$

where *Fc* is the Schroeder frequency, K is a constant (2,000 for SI units), RT is the reverberation time, and V is the volume. For instance, in a 5 m x 4 m x 3 m room with a 0.8 s reverberation time, *Fc* is close to middle C on the piano: 224 Hz. Below the Schroeder frequency, the individual characteristics of the room become apparent in the resonances seen in the frequency response at a given receiver position. Alternatively, the statistical nature of reflections above the Schroeder frequency makes the reverberation times indicated in various standards meaningful for describing the character of the room [59]. In other words, the pattern of early reflections and the relative damping of low-frequency modes becomes of especial significance to the individual character of the room at a given listening position. This will manifest timbral changes to program material that certainly affect audio quality, and can be assumed to carry over to audio-visual interaction.

3.3 Cross-modal compensation

Cross-modal mismatches in perceived quality are of interest because of the potential for "cross-modal compensation," i.e., the ability to create relative shifts in perceptual quality in one modality by changing another. For this reason, it is recognized that perceived quality in a multi-modal presentation cannot be assessed by examining the separate modes in isolation and then attempting to predict interactions [2]. A commonly made observation by those working in gaming applications is that "really high-quality audio will actually make people tell you that games have better pictures, but really good pictures will not make audio sound better; in fact, they make audio sound worse" [60]. A recently completed dissertation investigated the interaction of three different resolution levels of audio and visual displays [61]. A statistically significant effect was found showing that a visual display that would, without sound, be rated as "medium" quality would be elevated to "high" quality with the addition of either medium or high quality sound. A similar observation was made in another study with 367 non-experts at a shopping mall, who consistently evaluated television images with high fidelity or stereo sound as "more interesting, more involving and better liked" compared to low fidelity or monaural sound, although the powers of discrimination between audio variables was quite limited [62].

3.4 Effect of spatial displacement between audio and visual stimuli

When we can see a sound source, the auditory and visual localization cues are usually consistent and considered plausible. Under these conditions, the perceptual errors of problems of reversals and externalization can be virtually eliminated if the subject associates an acoustic stimulus with a visual source. Only audio localization can be used for sources outside the visual field-of-view, or when the sound source is hidden. In multimedia situations, the field-of-view reduces from ±90 degrees to about ± 50 degrees, depending on viewing distance. In home theater and related applications, there is a well-recognized potential for displaced location between auditory and visual stimuli, particularly involving the center and front left-front right speakers. This problem is aggravated with non-transparent screens when the center speaker for spoken dialogue has to be placed below or above the televised image.

The immediate perceptual response to a locational discrepancy between modalities is that the perceived location of the non-dominant modality will shift towards the other. This phenomenon is known as "intersensory bias." When visual and auditory cues conflict, sounds are localized to the position of the visual stimuli: this is known as the "ventriloquism effect" [63-65]. However, Perrott conducted a study where sequential and concurrent audio-visual stimuli presentations were varied, and concluded that visual dominance does not operate under all conditions of spatial presentation [4]. While the visual modality is dominant for stimuli positions at about 15-25 degrees azimuth, auditory information provides superior localization information at greater angles within the visual field. This suggests that "visual dominance" does not operate unilaterally across azimuth positions.

From a cognitive perspective, there is also evidence that the level of compellingness between spatially displaced audio and visual stimuli may be important. In other words, there is an influence caused by a subject's assumption that different modalities are indeed providing information about the same event. In one study, subjects listened to a voice while simultaneously viewing videotape of either a speaking person whose voice was the audio source, or of a light source whose brightness was modulated by the amplitude fluctuations of the voice. The visual and auditory stimuli were separated by 20 degrees. The voice and face appeared fused approximately 78% of the time, while the voice and the light were fused only 49% of the time [66].

Several studies have examined visual-auditory displacement in the context of future home theater technologies. Komiyama measured "annoyance" in terms of a 5 point scale ("imperceptible" to 'annoying") for both expert and non-expert subjects, using a visual image at 0 degrees and loudspeaker positions at 0, 5, 10,15, 20 30 45, 90 135 and 180 degrees [67]. The screen itself extended ±15 degrees. The results indicated a tolerance of 11 degrees for experts and 20 degrees for non-experts. Interestingly, back-front reversals occurred for the 135 and 180 positions and so subjective responses for these positions were eliminated from the data, presumably because the general trend between annoyance and speaker angle reversed itself at these locations. Theile proposed a "congruity index" to express the relationship between speaker displacement and screen extent, specifically, a listening angle/viewing angle ratio of 1.2 (e.g., speakers at 60 degrees correspond to a viewing angle of 50 degrees) [68]. Naturally, the value of this index changes in relationship to the viewer's distance. Woszczyk goes as far as to conclude that "… precision of directional match between sound and picture are not justified because vision alone dominates the localization of sources we see" [69] .

Bech gathered judgements of "visual impression" quality in terms of the match between the acoustical properties of the environmental context, using a broadcast from a small television with the acoustics of a cathedral as an example of the lowest rating [48]. This question is notable in that it asks subjects to match subjective impressions of environmental contexts, although the impression of the reverberation in the absence of the visual cue is not known, and therefore visual influence per se cannot be assessed. The results indicated that the quality of spatial impression reproduction was correlated with increasing speaker angle, and listening position was found to have a significant influence.

A final observation regarding spatial asynchrony of visual and audio is that most of the studies have been is concerned with static cues. Since dynamic cues present a greater challenge to the perceptual system than static ones, it is necessary to ascertain how the visual and auditory cues interact under these conditions, in order to calibrate the cues and determine the change in their interaction as a function of movement. Little research has been done in this area as of yet, although some preliminary work has been undertaken at NASA Ames in a virtual environment context.

3.5 Audio-visual temporal asynchrony

Wenzel has investigated asynchrony problem from the perspective of virtual acoustic displays, by manipulating the delay between a head tracking device and the acoustical rendering, at delays up to 500 ms [70]. Surprisingly, the effect on absolute localization was rather low, suggesting that there was subjective adaptation that compensated for the delay. The effect of asynchrony between simultaneously presented visual and audio virtual stimuli remains to be assessed.

Less pertinent to spatial reproduction but probably foremost in terms of overall quality is the effect of temporal asynchrony between audio and visual media on speech intelligibility. The fact that speech is more intelligible with the presence of lip of the speakers is evidenced in the procedure for calculation of the speech articulation index (AI – see [71]). The value of AI falls off much less steeply as a function of decreasing signal-noise when visual cues for lip reading are available. The effect of audio-visual interaction on speech intelligibility is well known from the McGurk effect, where intelligibility is affected by conflicting visual-audio cues [72].

The effect of multi-modal asynchrony was investigated Rimell, Hollier and Voeckler from the perspective of quality degradation [2]. Dixon and Spitz investigated the detectability of audio-visual asynchrony [73]. Subjects detected asynchronies at 257.9 ms delay and 131.1 ms advance for a speaking voice, and 187.5 ms delay and 74.8 ms advance for a film of a hammer hitting a peg. Results overall indicate almost unanimously that asynchrony is much easier to detect when sound precedes rather than follows an associated visual pattern.

3.6 Vibro-tactile and vibro-acoustic interaction

A common experience while driving a vehicle is to use multi-sensory feed back from tactile and auditory cues to monitor road conditions while navigating the automobile visually. The sense of the road condition (e.g., its smoothness) becomes diminished without tactile cues manifested by vibration throughout the seat and steering wheel and accompanying aural cues. Vibration becomes most noticeable and disturbing at very high levels of acceleration, e.g., from sonic booms, construction, or machinery; naturally, everyone knows of a sound system capable of classification as an aggressive weapon. Far more interesting and varied are the more subtle types of vibration that are present in homes, offices, on your desk near your workstation, in the windows of your home, and in the floor of your multimedia presentation room. There are often sources in the environment that activate both "structure-borne" vibration, where sound is transmitted via the walls, floors, ceiling and other elements of a room, and airborne-induced vibration that can cause structural elements to vibrate. These same sources can also act as airborne and structure-borne sound sources. Once a wall, desk or other object is set into motion via the vibration, it is then possible to have objects rattle and make noises of their own, such as picture frames hanging on a wall. Large surfaces are especially efficient radiators of sound; this is referred to in one source as the "sounding board effect" with reference to piano sounding boards [74]. Under certain conditions, the receiver can potentially hear, see, and feel the effect of vibration.

There is a large range of frequencies (roughly 20-150 Hz) that are potentially both audible and tactile, and could therefore be referred to as the "vibro-acoustic frequency range." However, most virtual acoustic and audio-visual simulations are of uncoupled spaces with perfect sound isolation (there are a couple of notable cinematic exceptions). Why then simulate vibration? The main reason is that by ignoring or not simulating its presence, the vibration in the real environment of the listener predominates and could potentially conflict with the intended audio-visual virtual experience. Commonly reported experiences of "feeling the music" in a live music situation are worth simulating in recorded playback, but there would be a conflict between the vibration of a ceiling-mounted air conditioner and the experience of a simulation of a 17th century Baroque music performance

Some applications emphasizing human performance have included vibration cues. Flight simulators capable of motion simulation provide inertial cues, but these do not transmit structure-borne vibration or vibro-acoustic cues in any realistic sense. Rather, the goal of the motion simulation is to provide proper correlation between vestibular and visual cues [75]. Furthermore, while the overall sound quality reacts to engine controls in a predictable manner, there is an easily recognized mismatch between the sound field in a simulator cockpit and that found in most simulators. Haptic interfaces have also received attention in the development of virtual reality systems, but with the emphasis on enabling manual interaction between an operator and the virtual environment. The predominant area of research involves the hand. The goal would be to allow a much more complex and naturalistic interaction between human and machine than is currently available with the familiar computer keyboard or mouse. Force feedback is also considered an important technology development for virtual reality applications [35].

The subject of vibration perception is highly variable as a function of context, sating position, and of the individual; only generalizations can be made. An international standard for building vibration generalizes that "Experience has shown in many countries that complaints regarding building vibrations…are likely to arise when the vibration levels are only slightly in excess of perception levels" [76]. Figure 2 shows vibration thresholds in conjunction with low end of the NC curves described earlier. This figure indicates how vibration is both "feelable" and "audible" when acceleration levels are relatively high. For instance, a wall vibrating with an acceleration level of .01 g. would be quite audible, but barely perceptible in terms of vibration (the 63 Hz octave band sound pressure level for NC 60 is equivalent to 77 dB). However, the same level of acceleration at infrasonic frequencies would be very perceptible. In spite of this sensitivity, humans probably disregard low-amplitude vibrations, allowing other modalities to dominate. This is analogous to the reason why people put up with low-quality loudspeakers in television sets for so many years, even when they owned a high-fidelity audio system for music.

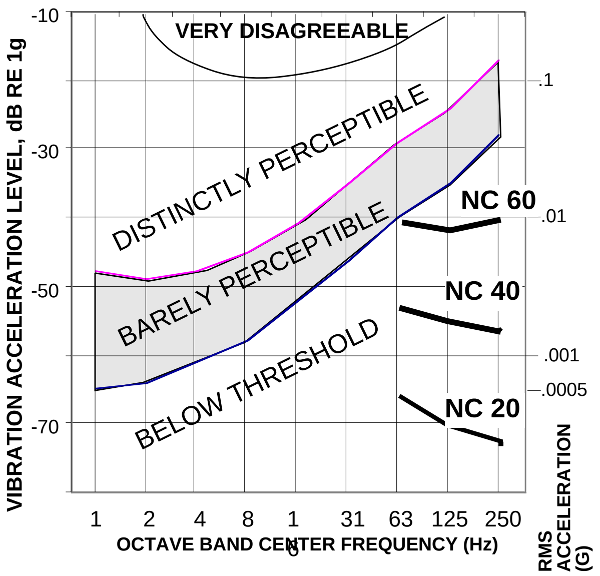


Figure 2. Approximate sensitivity and response to feelable vibration, with low-frequnecy end of the NC curves indicated (adapted from [74]).

The use of multi-modal simulation has been of interest for automotive sound quality assessment. Otto and his colleagues reported recently on the development of the “Ford Vehicle Vibration Simulator,’ a platform using binaural sound playback and a vehicle vibration simulator [77, 78]. The goal is to correlate subjective responses to the simulations to objective analysis of components and design of the automobile. The vibration simulator is unique in that it can simultaneously vibrate the seat with 6 degrees-of-freedom (DOF), the steering wheel with 4 DOF, and the floor and brake or accelerator pedal with 1 DOF. Audio simulations are based on binaural recordings and accelerometer measurements from test vehicles or devices. Eventually, they hope to be able to accomplish sound quality assessment on virtual designs. Their simulation techniques are motivated by the hypotheses that the presence or absence of sound influences what are usually assumed to be purely tactile percepts, and conversely that vibration influences the perception of sound. In other words, the interaction between modalities can make a particular component either more or less noticeable.

4. CONCLUSIONS

This paper has summarized some of the research for multi-modal interaction as it pertains to spatial perception, particularly for those application areas where 3-D sound manipulation is of interest. Some of the philosophical differences between localization studies and sound quality evaluations of spatial sound were reviewed. The influence and possible detrimental effects of room acoustics, vibration, listening position, and spatial and temporal asynchrony were reviewed, with an emphasis on some of the work reported in various publications of the Audio Engineering Society. It is hoped that some of the perspectives and observations presented here will be of use to future researchers who wish to pursue the challenges of assessing cross-modal interaction for future technology development.

ACKNOWLEDGEMENTS

Many thanks are due to my colleagues who helped in the review of this manuscript. This work was supported by NASA Ames Research Center-San José State University Foundation cooperative research grant NC2-327.

REFERENCES

[1] Welch, R.B. and D.H. Warren, "Intersensory interactions." In K.R. Boff, L. Kaufman, and J.P. Thomas (Eds.), *Handbook of perception and human performance*, New York: Wiley (1986).

[2] Rimell, A.N., M.P. Hollier, and R.M. Voelcker, "The influence of cross-modal interaction on audio-visual speech quality perception," *Proceedings of the Audio Engineering Society 105th Conference, San Francisco,* Preprint 4791 (1998).

[3] Hollier, M.P. and A.N. Rimell, "An experimental investigation into multi-modal synchronisation sensitivity for perceptual model development," *Proceedings of the Audio Engineering Society 105th Conference, San Francisco,* Preprint 4790 (1998).

[4] Perrott, D.R., "Auditory and visual localization: two modlaities, one world," *Audio Engineering Society 12th International Conference. The perception of reproduced sound.*, Copenhagen, pp. 221-231(1993).

[5] Woszczyk, W., S. Bech, and V. Hansen, "Interactions between audio-visual factors in a home theatre system: definition fo subjective attributes," *Audio Engineering Society 99th conference, New York,* Preprint 4133 (1995).

[6] Ellis, S.R., "Presence of Mind: a reaction to Sheridan's "Further musings on the psychophysics of telepresence"," *Presence*, Vol. 5, pp. 247-259, (1995).

[7] Begault, D.R., S.R. Ellis, and E.M. Wenzel, "Headphone and head-mounted visual displays for virtual environments," *Proceedings of the Audio Engineering Society 15th International Conference. Audio, Acoustics,*

and Small Spaces, Snekkersten, DK, pp. 213-217 (1998).

[8] Fisher, H. and S.J. Freedman, "The role of the pinnae in auditory localization," *Journal of Auditory Research,* Vol. 8, pp. 15-26, (1968).

[9] Wenzel, E.M., F.L. Wightman, D.J. Kistler, and S.H. Foster, "Acoustic origins of individual differences in sound localization behavior," Vol. 84, pp. S79, (1988).

[10] Wenzel, E.M., M. Arruda, D.J. Kistler, and F.L. Wightman, "Localization using non-individualized head-related transfer functions," *Journal of the Acoustical Society of America*, Vol. 94, pp. 111-123, (1993).

[11] Møller, H., M.F. Sørensen, C.B. Jensen, and D. Hammershøi, "Binaural technique: do we need individual recordings?," *Journal of the Audio Engineering Society*, Vol. 44, pp. 451-469, (1996).

[12] Kistler, D.J. and F.L. Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," *Journal of the Acoustical Society of America*, Vol. 91, pp. 1637-1647, (1992).

[13] Huopaniemi, J., N. Zacharov, and M. Karjalainen, "Objective and subjective evaluation of head-related transfer function filter design," *Audio Engineering Society 105th Convention, San Francisco,* Preprint 4805 (1998).

[14] Wightman, F.L. and D.J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *Journal of the Acoustical Society of America*, Vol. 91, pp. 1648-1661, (1992).

[15 Begault, D.R., "Virtual acoustics, aeronautics and communications," *Journal of the Audio Engineering Society*, Vol. 46, pp. 520-530, (1998).

[16] Perrott, D., J. Cisneros, R. McKinley, and W. D'Angleo, "Aurally Aided VIsual Search under Virtual and Free-Field Listening Conditions," *Human Factors*, Vol. 38, pp. 702-715, (1996).

[17] Bronkhorst, A.W., J.A. Veltman, and L. van Breda, "Application of a three-dimensional auditory display in a flight task," *Human Factors*, Vol. 38, pp. 23-33, (1996).

[18] Crispien, K. and T. Ehrenberg, "Evaluation of the "cocktail party effect" for mulitple speech stimuli within a spatial auditory display," *Audio Engineering Society 97th Convention, San Francisco,* Preprint no. 3891 (1994).

[19] Begault, D.R. and T. Erbe, "Multichannel spatial auditory display for speech communications," *Journal of the Audio Engineering Society*, Vol. 42, pp. 819-826, (1994).

[20] Begault, D.R., "Virtual acoustic displays for teleconferencing: Intelligibility advantage for "telephone grade" audio," *Audio Engineering Society 98th Convention, Paris,* preprint 4008 (1995).

[21] Ericson, M.A. and R.L. McKinley, "The intelligibility of multiple talkers separated spatially in noise." In R.H. Gilkey and T.R. Anderson (Eds.), *Binaural and Spatial Hearing in Real and Virtual Environments*, pp.701-724. Mahwah, New Jersey: Lawrence Erlbaum Associates (1997).

[22] Bronkhorst, A.W. and R. Plomp, "Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing," *Journal of the Acoustical Society of America*, Vol. 92, pp. 3132-3139, (1992).

[23] Thurlow, W.R., J.W. Mangels, and P.S. Runge, "Head movements during sound localization," *Journal of the Acoustical Society of America,* Vol. 42, pp. 489-493, (1967).

[24] Thurlow, W.R. and P.S. Runge, "Effects of induced head movements on localization of direct sound," *Journal of the Acoustical Society of America*, Vol. 42, pp. 480-487, (1967).

[25] Wallach, H., "The role of head movements and vestibular and visual cues in sound localization," *Journal of Experimental Psychology*, Vol. 27, pp. 339-368, (1940).

[26] Begault, D.R., "Perceptual effects of synthetic reverberation on three-dimensional audio systems," *Journal of the Audio Engineering Society*, Vol. 40, pp. 895-904, (1992).

[27] Durlach, N.I., A. Rigopulos, X.D. Pang, W.S. Woods, A. Kulkkarni, H.S. Colburn, and E. Wenzel, "On the externalization of auditory images," *Presence: Teleoperators and Virtual Environments*, Vol. 1, pp. (1992).

[28] Griesinger, D., "Binaural techniques for music reproduction," *Proceedings of the AES 8th International Conference*, Washington, D.C., pp. 197-207(1990).

[29] Shinn-Cunningham, B.G., N.I. Durlach, and R.M. Held, "Adapting to supernormal auditory localization cues. I. Bias and resolution," *Journal of the Acoustical Society of America*, Vol. 103, pp. 2656-2666, (1998).

[30] Durlach, N.I., B.G. Shinn-Cuningham, and R. Held, "Supernormal auditory localization. I. General background.," *Presence*, Vol. 2, pp. 89-103, (1993).

[31] Blauert, J.,*Spatial hearing: The psychophysics of human sound localization.* Cambridge: MIT Press (1983).

[32] Begault, D.R. and E.M. Wenzel, "Headphone Localization of Speech," *Human Factors*, Vol. 35, pp. 361-376, (1993).

[33] Wightman, F.L. and D.J. Kistler, "Headphone simulation of free-field listening. II: Psychophysical validation," *Journal of the Acoustical Society of America*, Vol. 85, pp. 868-878, (1989).

[34] Wenzel, E.M., "What perception implies about implementation of interactive virtual acoustic environments," *Audio Engineering Society 101st Convention*, Los Angeles, preprint 4353 (1996).

[35] Begault, D.R.,*3-D Sound for Virtual Reality and Multimedia.* Cambridge, MA: Academic Press Professional (1994).

[36] Bech, S., "Perception of reproduced sound: Audibility of individual reflections in a complete sound field, II," *Audio Engineering Society 99th Convention*, preprint 4093 (1995).

[37] Olive, S.E. and F.E. Toole, "The detection of reflections in typical rooms," *Journal of the Audio Engineering Society*, Vol. 37, pp. 539-553, (1989).

[38] Begault, D.R., "Audible and inaudible early reflections: thresholds for auralization system design," *Audio Engineering Society 100th Convention*, preprint no. 4244), Copenhagen, DK, (1996).

[39] Haas, H., "The influence of a single echo on the audibility of speech," Vol. 20, pp. 146-159, (1972).

[40] Begault, D.R., "Binaural Auralization and Perceptual Veridicality," *Audio Engineering Society 93rd Convention*, San Francisco, preprint 3421 (1992).

[41] Rumsey, F., "Subjective assessment of the spatial attributes of reproduced sound," *Proceedings of the Audio Engineering Society 15th International Conference (Audio, Acoustics, and Small Spaces), Snekkersten, DK*, pp. 122-135(1998).

[42] ITU, *Methods for the subjective assessment of sound quality-general requirements*, ITU-R BS.1284. Geneva: International Telecommunications Union (1997).

[43] Toole, F.E., "Subjective measurements of loudspeaker sound qulaity and listener performance," *Jounral of the Audio Engineering Society*, Vol. 33, pp. 2-32, (1984).

[44] Miyasaka, E., "Methods of quality assessment of multichannel sound systems," *Proceedings of the Audio Engineering Society 12th International Conference,* Copenhagen, pp. 188-196 (1993).

[45] ITU, *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems,* ITU-R BS.1116-1. Geneva: International Telecommunications Union (1994-1997).

[46] ITU, *Methods for the subjective assessment of audio systems with accompanying picture,* ITU-R BS.1286. Geneva: International Telecommunication Union (1997).

[47] Beranek, L.,*Concert and opera halls. How they sound.* Woodbury, NY: Acoustical Society of America (1996).

[48] Bech, S., "The influence of stereophonic width on the perceived quality of an audiovisual presentation using a multichannel sound system," *Journal of the Audio Engineering Society*, Vol. 46, pp. 314-322, (1998).

[49] Griesinger, D., "Multichannel sound systems and their interaction with the room," *Audio Engineering Society 15th International Conference. Audio, Acoustics and Small Spaces,* Snekkersten, DK, 159-173(1998).

[50] Zacharov, N., "Subjective appraisal of loudspeaker directivity for multichannel reproduction," *Journal of the Audio Engineering Society*, Vol. 46, pp. 288-303, (1998).

[51] Begault, D.R., "Challenges to the successful implementation of 3-D sound," *Journal of the Audio Engineering Society*, Vol. 39, pp. 864-870, (1991).

[52] Holman, T., "Scaling the experience," *Audio Engineering Society 12th international Conference. The perception of reproduced sound,* Copenhagen, pp. 232-251(1993).

[53] Altman, R., "Sound Space." In R. Altman (Ed.), *Sound Theory/Sound Practice*, New York: Routledge (1992).

[54] Amet, E.H., *Method of and means for localizing sound reproduction,* United States of America Patent 1,124,580 (1915).

[55] Maxfield, J.P., "Some physical factors affecting the illusion in sound motion pictures," *Journal of the Acoustical Society of America*, Vol. 3, pp. 69-80, (1931).

[56] Toole, F.E. and S.E. Olive, "Hearing is believing vs. believing is hearing: blind vs. sighted listening tests, and other interesting things," *Audio Engineering Society 97th Conference*, San Francisco, Preprint 3894. (1994).

[57] Harris, C.M., *Handbook of Acosutical Measurements and Noise Control (3rd Ed.)*. New York: McGraw-Hill (1991).

[58] Schuck, P.L., S.E. Olive, J.G. Ryan, F.E. Toole, S.L. Sally, M.E. Bonneville, K.L. Momtahan, and E.S. Verreault, "Perception of perceived sound in rooms: some results of the athena project," *Audio Engineering Society 12th International Conference. The perception of reproduced sound.,* Copenhagen, 49-73(1993).

[59] Kuttruff, H., "Sound fields in small rooms," *Audio Engineering Society 16th international conference. Audio, acoustics and small spaces.,* pp. 11-15 (1998).

[60] Tierney, J., *Jung in Motion, Virtually, and Other Computer Fuzz,* in *The New York Times*. 1993: New York CIty. p. B1-4.

[61] Storms, R.L., *Auditory-visual cross-modal perception phenomena* (unpublished dissertation), Naval Postgraduate School, Monterey, California (1998).

[62] Neuman, W.R., A.N. Crigler, and M.V. Bove, "Television sound and viewer perceptions," *Audio Engineering Society 9th International Conference. Television sound today and tomorrow.,* Detroit, pp. 101-104 (1991).

[63] Thurlow, W.R. and C.E. Jack, "Certain determinants of the "ventriloquism effect"," *Perceptual and Motor Skills*, Vol. 36, pp. 1171-1181, (1973).

[64] Howard, I.P. and W.B. Templeton, *Human Spatial Orientation.* New York: Wiley (1966).

[65] Posner, M. and M. Nissen, "Visual Dominance: an information-processing account of its origins and significance," *Psychological Review*, Vol. 83, pp. 157-171, (1976).

[66] Radeau, M. and P. Bertelson, "Adaptation to auditory-visual discordance and ventriloquism in semirealistic situations," *Perception and Psychophysics*, Vol. 22, pp. 137-146, (1977).

[67] Komiyama, S., "Subjective evaluation of angular displacement between picture and sound directions for HDTV sound systems," *Journal of the Audio Engineering Society*, Vol. 37, pp. 210-214, (1989).

[68] Theile, G., "Trends and activities in the development of multichannel sound systems," *Audio Engineering Society 12th International Conference. The perception of reproduced sound.,* Copenhagen, pp. 180-187 (1993).

[69] Woszczyk, W., "Quality assessment of Multichannel Sound Recordings," *Audio Engineering Society 12th International Conference. The perception of reproduced sound.,* Copenhagen, 197-218(1993).

[70] Wenzel, E.M., "Effect of increasing system latency on localization of virtual sources," *Audio Engineering Society 16th International Conference on Spatial Sound Reproduction* (in this volume) (1999).

[71] ANSI, *American National Standard Methods for the Calculatonof the Articulation Index*, ANSI S3.5-1969. New York: American National Standards Institute (1969).

[72] McGurk, H. and J. McDonald, "Hearing lips and seeing voices," *Nature*, Vol. 264, pp. 746-748, (1976).

[73] Dixon, N.F. and L. Spitz, "The detection of auditory visual desynchrony," *Perception*, Vol. 9, pp. (1980).

[74] Miller, L.N., *Noise Control for Buildings and Manufacturing Plants,* Cambridge, MA: Bolt, Beranek and Newman (1981).

[75] Ellis, S.R., "Nature and origins of virtual environments: a bibliographic essay," *Computing Systems in Engineering*, Vol. 2, pp. 321-347, (1991).

[76] ISO, *Evaluation of human exposure to whole-body vibration. Part 2: Continuous and shock-induced vibration in buildings,* 2631-2. Geneva: International Standards Organization (1989).

[77] Meier, R.C., N.C. Otto, and W.J. Pielemeir, "The Ford Vehicle Vibration Simulator for Subjective Testing," *Sound and Vibration*, Vol. 32, pp. 26-32, (1998).

[78] Otto, N.C., B.J. Feng, and G.H. Wakefield, "Sound Quality Research at Ford- Past, Present and Future," *Sound and Vibration*, Vol. 32, pp. 20-23, (1998).