ICAT 2001
December 5-7, Tokyo, Japan
THE VIRTUAL REALITY SOCIETY OF JAPAN

Binaural range display with independent control over loudness: Presenting nearby whispers in the dark

William L. Martens

Multimedia Systems Lab, University of Aizu
Aizu-Wakamatsu 965-8580, Japan
voice: [+81](242)37-2791; fax: [+81](242)37-2731
e-mail: wlm@u-aizu.ac.jp

Abstract

A prototype bimodal display was implemented that separates the display space into two distinct regions – one for presentation of binocular visual information, the other for presentation of binaural auditory information. In this prototype, stereographic visual imagery is displayed over a 150° angle on the horizontal meridian via three large rear-projection screens, while the audio extends over a complementary 210°. The user of this visual display (which can be public or private) receives a personal presentation of virtual sound sources very near the user's head, but from behind, outside of the user's field of view. As the visual and auditory display spaces are spatially distinct, the delivery of multiple voice messages from locations outside of the frontal field of view constitutes a separate communication channel that will not complete for the user's limited capacities based upon spatial discrimination. This paper focuses upon the development of an effective binaural display technology employing a simplified model of range-dependence in simulated head-related transfer functions for headphone display of virtual sources at close range (within the user's personal space). The presentation of whispered speech originating from within the blind area just behind the user has been termed the "dark voice channel." User tests confirmed that the prototype auditory display supports the useful feature of varying the auditory range of equally loud nearby whispered speech. Based upon psychophysical functions for auditory range derived under conditions of expected use of the display, a calibrated deployment of the range control was accomplished. In contrast to most binaural display technology, then, this prototype system supports independent control over source loudness and range.
Keywords: human factors, auditory displays, spatial sound reproduction, binaural hearing, telepresence

1. Introduction

Just as the telephone brought explosive growth in global human communication, **I**nternet **P**rotocol (IP)-based services promise yet another revolution in interpersonal communication within the network-rich "global village" inhabited by human users of such technology. The rapid growth of internet use by networked devices other than general-purpose computers is being driven by mobile telephone services, such as i-MODE access, but a parallel growth of the technology to carry "**V**oice **o**ver **IP**" (VoIP) has begun to have a dramatic impact upon the telephony service market. Based upon the assumption that head-mounted personal audio systems will gain a substantial share of the market currently dominated by the hand-held mobile phone, advanced audio telecommunication features based upon binaural technology are likely to become more common [1]. This paper proposes to provide nearby localization of multiple voice messages within the portion of space outside the user's field of view. The delivery of voice messages from locations outside of the frontal field of view can provide a separate communication channel that minimizes interference between messages arriving at similar spatial incidence angles. This selective attention advantage is based upon spatial segregation between the visual display of messages within a frontal region and the auditory display of voice messages from the remaining "dark " region of space (which in a previous presentation [2] was termed the "dark voice channel").

When auditory and visual displays are integrated into a spatially coordinated human-computer interface, a first step in setting up such a multi-modal display system is to make sure that corresponding points in the respective perceptual spaces are in good registration to each other. The most direct approach to solving this problem is to execute an egocentric cross-modal matching task for the two display modal-

ities. In the current application, egocentric cross-modal matching was not appropriate, since all the auditory display technology under test was designed to deliver speech messages from outside the user's field of view, but at very close range. Furthermore, since whispered (unvoiced) speech contains relatively less reliable information about auditory source range than voiced speech [3], this investigation focussed upon human perception of the range of a whispering source (hence the subtitle, "Presenting nearby whispers in the dark"). In the studies reported here, subjective ratings of auditory range were made for a set of spatialized speech sounds relative to a fixed standard stimulus, seemingly located just outside the user's reach. The ranges of the variable comparison stimuli were specified to be closer than that of the standard stimulus, so that numerical estimates could be made relative to a constant, most distant reference. Furthermore, ratings of the perceived direction and spatial extent of the whispered speech stimuli were collected in order to form a more complete picture of the dimensions of the auditory spatial imagery associated with the display technology under test.

The reason user testing was required here stems from a fundamental difficulty in auditory range display, which is especially profound when spatial auditory display is non-interactive. In particular, interactive exploration of a virtual acoustic environment that employs a head-tracking auditory spatial display is most effective in ameliorating difficulties in determining the range of a virtual sound source. Of course, directional perception is also strongly affected by interactive auditory display in which changes in auditory image lateralization are coupled with voluntary rotations of the listener's head [4]. In the following sections, then, this paper begins by introducing some of the background that should aid in the reader's understanding of the importance of the current developments in auditory spatial display technology.

1.1 Visual Field and Auditory Space

The normal human visual field extends over a horizontal angle subtending more than 180° [5], though the binocular portion of this field typically is only about 120°.[1] When a visual display surface is positioned at a distance of about 5 m from the user, and the user's gaze is fixed upon the plane of that display surface, users typically have no difficulty with binocular fusion of disparate images of virtual objects positioned beyond the screen. In fact, positive binocular visual disparities as great as 200 arcmin will produce stereoscopic depth percepts that can be viewed without discomfort by most users. In contrast, for visual objects that are intended to appear much nearer to the user than screen distance, negative binocular visual disparities exceeding 200 arcmin will produce double images (no binocular fusion) and most viewers will begin to experience eyestrain. (For a more detailed treatment of optimal stereographic display of binocular disparities, see [7].)

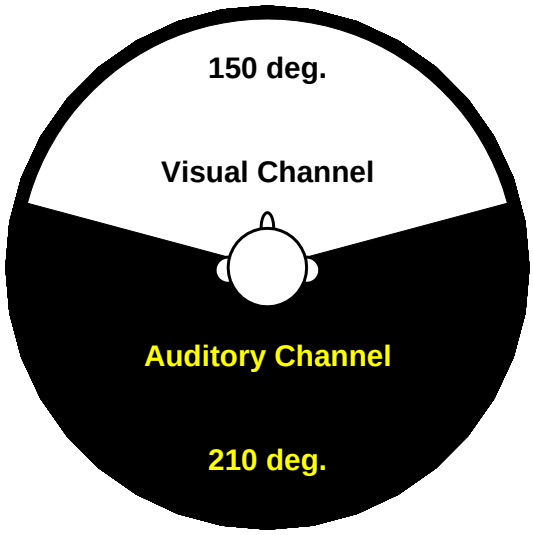

Fig. 1: Division of the display space surrounding the observer into a frontal 150° binocular visual display region and a complementary 210° binaural auditory display region.

When auditory information is presented via headphones using spatial audio signal processing that simulates the acoustics of the head (via head-related transfer functions), virtual sound sources can be effectively displayed over the 210° angle that complements the extent of the currently employed 150° binocular visual display. (see Figure 1). Furthermore, virtual sound sources can approach very near the user without producing any undesirable consequence akin to the eyestrain associated with nearby visual images. There is a psychological consequence of hearing sources at close range, within the user's personal space, but this consequence is not objectionable. In fact it has the desirable property of attracting the attention of the user when this intrusion is warranted for an important voice message. It is a great advantage for an auditory display to be able to reach into the user's personal space without increasing the loudness of the virtual source, so that other messages are not masked by large loudness differences. Of course, it is also a great advantage to control auditory range without changes in loudness, since the sound pressure level (SPL) of the source at the source position (as opposed

[1] Loss in visual field extent can result from optic nerve damage or diseases such as glaucoma, and can include the loss of binocular visual information if there is damage at the optic chiasm. Though there are considerable individual differences in visual field extent, it is interesting to note that the standard minimal requirements for safe operation of a private vehicle establish a norm for navigation. For example, the American Automobile Association (AAA) has established that an adequate visual field should be defined as 90° on the horizontal meridian, 45° to both the right and left, and 20° on the vertical meridian both above and below fixation [6].

to the listening position) can confound the listener's ability to determine whether changes in loudness occur due to changing intensity of changing range.

The idea here is simply to provide auditory information via a display channel that is spatially distinct from the visual channel. As the visual information display channel is typically located directly ahead of the user (front and center), that spatial region can become saturated with display elements, especially in avionics applications [8] [9]. Even when the visual display extends throughout the user's entire visual field (using such technology as the Vision Dome[2]), most of its critical information will be presented near focal vision. Of course, visual communication can be enhanced by coordinated auditory display of spatially co-located sound events, but the display within a shared spatial region of auditory information that is independent of the displayed visual information will often compete for the user's attention. Of course, it would use human spatial perceptual capacities better if the natural selective attention supported by human spatial hearing were exploited.

It may well be asked, however, why the display space should be separated between visual and auditory modalities. Though it is clear that the visual display will necessarily cover only the frontal region within which we have visual sensitivity, it may not be clear why the auditory display should be limited to the complementary spatial region. The answer to this question is two-fold. First, it should be noted that the proposed system is intended only for auditory displays that do not support updates of display parameters based upon realtime head-tracking data; therefore, observer attention cannot be sonically directed to visual targets outside of the visual field for active orientation behavior. Second, it is simply the case that headphone-based spatial auditory display technology does not position virtual sound sources very well within the frontal spatial region. Many headphone users presented with a high-fidelity binaural reproduction of a frontally located source will report that they hear the virtual source arriving from the rear rather than from the front [10] [11]. In contrast to such front $\rightarrow$ rear reversal, the opposite mismatch between recording location and reported location is less frequently observed. (In a related speech sound localization study, 30% of listeners commonly made rear $\rightarrow$ front reversals, while 70% made front $\rightarrow$ rear reversals [12]). A generalization of these results that is well-founded in the spatial hearing literature is the following: Auditory display of virtual sound sources within the rear hemifield is typically more successful than that within the front hemifield. Since this has been well documented [12] [4], and is so readily demonstrated [13], it seemed prudent to develop a spatial auditory display intended to present sounds only within the rear hemifield, in which success is so much more likely. It is also likely that less conscious attention would be required of users if they were to be informed of the fact that virtual sound sources would be presented only within the rear hemifield.

Human localization of transient visual events displayed anywhere within the the visual field can be practically effortless. Though visual acuity falls off rapidly outside of the fovea, only at the peripheral extremes of the visual field, which extend to the edges of the frontal hemifield, is significant increase in transient stimulation required for the detection and localization of a displayed visual event. In contrast, auditory spatial localization requires more conscious effort on the part of the user, and may require voluntary exploratory behavior to resolve common spatial ambiguities such as the distinction between front- and rear-ward incidence. Only left/right identification of sound source location is truly effortless for humans not engaged in active localization [14]. Auditory displays that do not respond to changes in the position or orientation of the user's head provide unreliable cues to front- and rear-ward incidence, and their use has revealed systematic biases in human spatial localization performance using such non-interactive auditory displays.

Besides these difficulties in control of the direction of virtual sound sources, modulation of the range (egocentric distance) of virtual sound sources is also somewhat problematic. As the ultimate goal here was the deployment of auditory display technology for which specified responses are in some way calibrated to the actual responses of the human listener, user tests of the implemented display technology were executed. Whereas the primary motivation in spatial hearing research has been to gain greater understanding of the mechanisms of human spatial hearing, the motivation for this applied research has been the verification and validation of practical (efficient) virtual acoustic rendering. Therefore, subjective ratings of auditory range were made for a set of spatialized speech sounds relative to a fixed standard stimulus, seemingly located just outside the user's reach. The ranges of the variable comparison stimuli was specified to be closer than that of the standard stimulus, so that numerical estimates could be made relative to a constant, most distant reference. User testing of this sort typically provide more reliable data than absolute range judgments, which are referenced to an internal standard unique to each individual even when the response requested uses physical units such as meters.

1.2 Level-Based Range Control: A Fundamental Ambiguity

The typical solution to the problem of controlling auditory range has been to use changing source loud-

[2] The Vision Dome is a personal display system that projects video onto a hemispherical surface, covering a wide visual angle on both the horizontal and vertical meridian. The Vision Dome is available from Elumens Corporation, both in monoscopic and stereoscopic models (see `http://www.virtual-reality.com/`).

ness as the primary cue, but this solution presents an ambiguous stimulus to the user as changes in SPL at the source position are not as easily discriminated as changes in loudness at the listening position. The ambiguity stems from the difficulty of determining whether changes in the *proximal* SPL (i.e., level at the ear) are caused by changes in *distal* SPL (i.e., level at the source position) or by changes in source range. In fact, it is conceivable that a change in level due to a change in source range could be accompanied by a complementary change in level at the source position (*distal* SPL), and these changes would go unnoticed as there would be no net change in level at the ear (*proximal* SPL).

The primary means available for resolving such ambiguity in auditory spatial display of source range is the inclusion of appropriate indirect sound, but there is another means available for resolving level-based ambiguity in the auditory range of virtual sources located at close range. For headphone-based auditory display, a modification of the HRTF deployment can also be implemented to allow for range-based transformation of the direct sound. Such range-based variation in the HRTF has been well-documented for sources within arm's reach of the listener [15], and their effectiveness in producing changes in range perception in the listener's "personal space" has been established [16]. This paper summarizes recent efforts to establish control of auditory range that is separable from control of *proximal* SPL, and hence enables the novel feature of localizing louder virtual sound sources at greater ranges than softer virtual sources.

1.3 Range Control for Nearby Sources

Typically, personal, headphone-based auditory spatial display technology either fails to project (externalize) the auditory image of the *distal* stimulus to a location in the listener's auditory representation of the surrounding space (i.e., no externalization means only "in-head localization" [17]), or the source may be well externalized, but projected to a location at some greater distance from the listener, most often via the inclusion of a significant amount of reverberation that is easily detectable by the listener. The goal of the experiment reported here, succinctly put, was to test an efficient means to place an externalized virtual sound source so close to the listener's ear that it enters the listener's "personal space" [18]. When the *distally* projected auditory image of a virtual sound source enters the listener's "personal space," a psychological boundary is crossed that potentially carries special meaning to users in particular applications such as telecommunication within shared virtual acoustic environments. If such an audio transformation were properly engineered (both perceptually valid and perceptually calibrated), a spoken message could be made to sound as if it were whispered into the ear of the recipient, letting them know, for instance, that the message was intended for them in confidence (providing what has been termed a "whisper function" [19]).

One of the continuing problems of headphone-based virtual acoustic imagery has been the difficulty of creating auditory images that are clearly outside of the listener's head using a minimum amount of audio signal processing. The use of HRTFs is conventionally regarded as the first step in creating externalized auditory imagery [20], but this is only moderately effective for some spatial directions of the sound source [21], and truly unreliable [22]. Without simulated indirect sound of some sort, the likely result in headphone listening is "in-head localization" of the auditory image of the reproduced sound source [17]. The desired result in headphone reproduction, in contrast, is "out-of-head localization" and has often been shown to be dependent upon the amount of reverberation present in sound reproduction (c.f., [23]). Without reverberation, externalization is usually not experienced for sources arriving from incidence angles near the listener's median plane. But sources arriving from angles well removed from the median plane, especially those nearer to the listener's head, are much more commonly externalized.

Of course there are many acoustical sources of information cueing auditory range that are available to the human listener at close range, but this study focused only on the very simple case of dry HRTF-based processing (i.e., involving no indirect sound simulation). In particular, it was the level-based cues contained in the direct sound that seemed worth examining. The idea that variation of interaural level difference (ILD) in the direct sound might aid the listener in detecting range of nearby sources is not new (the hypothesis was probably first stated clearly by Hartley and Fry in 1922 [24]). Brungart [25] confirmed that listeners indeed do better in localizing nearby anechoic sources when those sources are well lateralized away from the listener's median plane. Of course, there is also a low-frequency boost for sources located near the head, which von Békésy pointed out in a 1938 paper [26] could potentially aid the listener in determining auditory range. However, in the author's recent related study [16] that compared the relative salience of four sources of acoustical information associated with range perception for the human listener, the level-based features of the direct sound (ipsilateral-ear SPL and ILD) dominated other cues (indirect-to-direct sound ratio and high-frequency attenuation of a filter simulating the head-shadow). Also, adding range variation to dry HRTF-based processing with simple gain adjustment requires less computation than indirect sound synthesis [27]. Indeed, it is quite difficult to efficiently generate effective simulations of indirect, or reverberant, sound, and inexpensive computational models often lead to poor results [28].

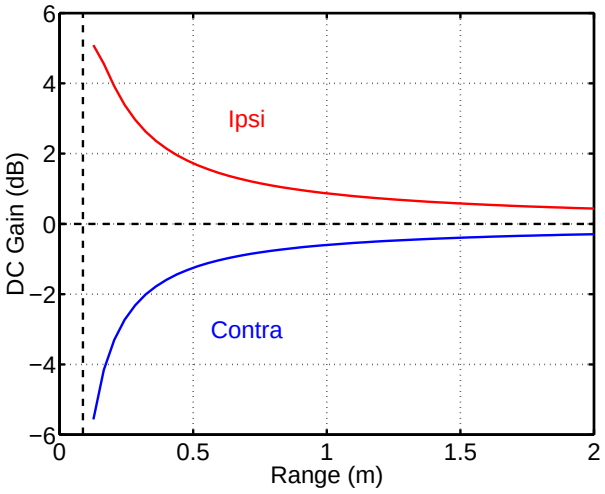

Fig. 2: Free-field normalized gain at DC for a point on the surface of an ideal spherical "head" as a function of source range from the center of the sphere. The upper curve (labeled "Ipsi") corresponds to the ipsilateral "ear" response for a source incidence angle of 130° azimuth, while the lower curve (labeled "Contra") corresponds to the contralateral response at that angle. The vertical dashed line shows the model "head" radius of .0875 m.

2. Methods

2.1 Stimuli

The choice to focus on ILD-based range cues via a simple manipulation of ipsilateral and contralateral SPL in this listening experiment was motivated by the lack of an efficient yet effective technology for creating auditory images that are clearly outside of the listener's head, but nonetheless very close to the listener's ear. Free-field listening in an anechoic chamber confirms that sources within around 1 m of the listener's head are often externalized and have a special quality that tells the listener that the source is nearby. Although measured HRTFs show a rather complex dependence on range, an analysis of the range dependence in the response of an ideal spherical receiver shows one very striking feature that might account for this special quality. That feature is shown in Fig. 2 for a source at 130° azimuth (from calculations based upon the model reported in [29]). Though the free-field normalized gain at DC is nearly 0 dB for sources arriving from ranges greater than 2 m, the interaural level difference (ILD) at DC gain grows large as the source approaches the listener's ear. As the gain increases at the listener's ipsilateral ear, the gain decreases at the listener's contralateral ear. In contrast to the relative auditory range cue provided by loudness, it has been hypothesized that this range-dependent variation in ILD might provide a more absolute range cue to the listener for nearby sources, and that auditory images might be effectively externalized for sources well removed from the listener's median plane even though indirect sound is absent from the simulation.[3]

For this experiment, three short utterances, "ha," "hi," and "hu," whispered by three Japanese talkers, were recorded at a range of 1 m. in a large anechoic chamber at the University of Aizu. Whispered (unvoiced) speech was chosen rather voiced speech, since whispered speech contains relatively less reliable information about actual *distal* SPL than voiced speech [3]. Likely due to the recognition of "vocal effort" associated with particular speech production levels, voiced speech contains range cues that seem to be inherent in the timbre of the sound source itself [31]. The three vowel sounds whispered by the talkers for this study span the vowel space defined by the first two formant frequencies of the vocal tract, and they represent the extremes of vowel coloration in the Japanese language.[4] The intention of this sound source selection was to allow the spatially-processed stimuli to vary in timbre as widely as possible while maintaining the same aspiration /h/ in the consonant-vowel (CV) stimuli. The transient, high-frequency content of the /h/ consonant was included to provide adequate stimulation for perceptual fusion of the stimulus into a single, coherent auditory image of the whispered speech sound.

The spatial sound processing of the stimuli was a variation of the conventional convolution with a pair of HRTFs (subject MES, see [33]) with no simulated indirect sound. The baseline convolution condition used an anechoically measured HRTF pair for a source arriving from the rear at 130° azimuth, at 0° elevation, and from a range of 1.5 m. Auditory range was then manipulated strictly in terms of the relative SPL of the ipsilateral- and contralateral-ear signals. The level of the resulting ipsilateral-ear signal was increased from its baseline level in three 3 dB steps (gain was 0 dB, 3 dB, 6 dB, and 9 dB). The level of the resulting contralateral-ear signal was decreased from its baseline level in three 3 dB steps (gain was 0 dB, -3 dB, -6 dB, and -9 dB). No frequency-dependent component of the head-response model was manipulated in order to match the change in the head shadow as sources approach ranges closer than 1 m (c.f., [16]).

[3] It should be noted that the modulation of ILD with range is most pronounced at low frequencies, and that therefore a more accurate simulation would model the frequency-dependence of this effect of range on the HRTF. Though simple filtering models for this effect have been implemented by the author and others (e.g., [30]), the interest in the present study was to test an extremely simple model based only upon frequency-independent gain adjustments, avoiding the variation in tone color that results from more accurate simulations.

[4] On average, the first formant frequencies range from 280 to 750 Hz and the second formant frequencies range from 1100 to 2300 Hz when Japanese subjects read word lists [32]. For the speech stimuli used in this study, the first two formant frequencies averaged over four talkers (based on LPC spectra of the vowel segments of their recorded speech) were approximately the following: /a/ – 750, 1200, /i/ – 280, 2280, and /u/ – 310, 1220 Hz.

A factorial combination of ipsilateral and contralateral level was executed by crossing the four possible values for each stimulus, and each of 9 sound sources (3×3 factorial of talkers and vowels) was combined in a randomized order for these 16 SPL combinations (4×4 factorial of ipsi and contra), to produce the set of 144 stimuli comprising a single experimental listening session (of which three were completed by each subject).

2.2 Subjects

Five subjects voluntarily participated in this experiment, four of whom were students at the University of Aizu. One was the author, a researcher with a substantial history of participation in similar listening tests. All were audiologically normal, with no reported hearing loss.

2.3 Procedure

The stimuli were presented to the listeners via Sennheiser HD590 headphones using the standard audio conversion hardware of an SGI workstation. The listening test was completed in three blocks of 144 trials each. In each trial, two stimuli were presented with a one-second inter-stimulus interval: a comparison stimulus of variable range, and a standard stimulus of fixed range. The baseline convolution condition (no gain adjustment) served as the standard stimulus that provided a reference by which listeners judged the range of the other experimental stimuli. After hearing first the standard and then the comparison stimulus, the listener was asked to rate the range of the comparison on a scale from 0 to 10. The value of 0 was to be reported if the auditory image was located inside of the listener's head. The value of 1 was to be used if the sound source seemed to be located extremely close to the listener's ear (i.e., the "verged cranial" position). The other extreme of the scale was anchored to the perceived range of the standard stimulus, and the response of 10 was to be given to any source that was perceived at roughly the same range as the that of the standard stimulus. The value of 9 was to be given to a source just noticeably closer than the standard.

A three-minute training session was completed before beginning the experimental trials, during which time listeners were to attempt to establish criteria for their use of the rating scale. The inter-trial interval during the training session was just one second, and this interval was increased to five seconds during experimental trials to allow time for a response to be generated.

3. Results

The influence of ipsilateral gain and contralateral attenuation on mean range ratings for one subject are shown in Fig. 3. The highest mean range ratings were obtained for a comparison stimulus that was identical to the standard stimulus. Thus, of all the combinations of ipsilateral and contralateral SPL present in the set of comparison stimuli, the greatest source range was reported when the nine sound sources were convolved with the unmodified HRTF (the standard stimulus, and the baseline convolution condition). As the ipsilateral gain was increased, and contralateral attenuation was held constant at 0 dB, reported range decreased for the source, which consistently seemed to be arriving from the rear at around 130° azimuth.

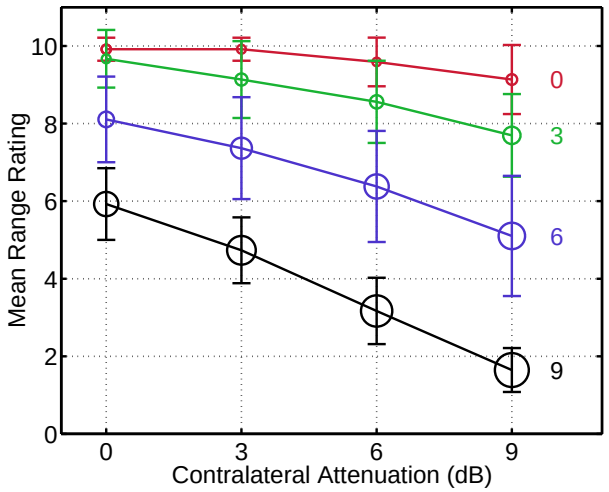

Fig. 3: Grand mean range ratings for one subject plotted over contralateral attenuation levels, combining results for three vowel sounds whispered by three native speakers of Japanese. The parameter of the graph is ipsilateral gain level, the values of which are used to label each of the four sets of connected plotting symbols (numerals appearing just to the right of each). Each symbol includes a standard error bar for each respective mean. The size of the symbols correspond to the grand mean rating of spatial extent for the auditory image associated with each stimulus.

The rightmost symbols in the graph shown in Fig. 3 show this decrease from range ratings of around 10 to ratings around 6. The plotting symbols grow larger in the graph as ipsilateral gain is increased, as a reminder that these sources are approaching the listener (i.e., simulating visual looming). When the ipsilateral gain was held constant at 0 dB, and contralateral attenuation was increased, reported range also decreased, but not to such a great extent. The topmost curve, labeled "0" in the plot, shows decreases from range ratings of around 10 to ratings of only around 9. But when the ipsilateral gain was held constant at an SPL 9 dB greater than that of the standard stimulus, increasing contralateral attenuation produced much lower range ratings (decreasing from ratings of around 6 to ratings below 2). It is clear from these data that larger ILDs result in increasingly closer source localization as the ipsilateral SPL increases. In effect, the close-range ILD cue works best when the whispered speech is so loud that it is already likely to produce

a lower range rating. Nonetheless, an analysis of the obtained range data enables independent control over loudness and range, by properly traversing the control surface defined by the two parameters, ipsilateral gain and contralateral attenuation. The mechanism whereby this control surface, illustrated in Fig. 4, is via inverse prediction.

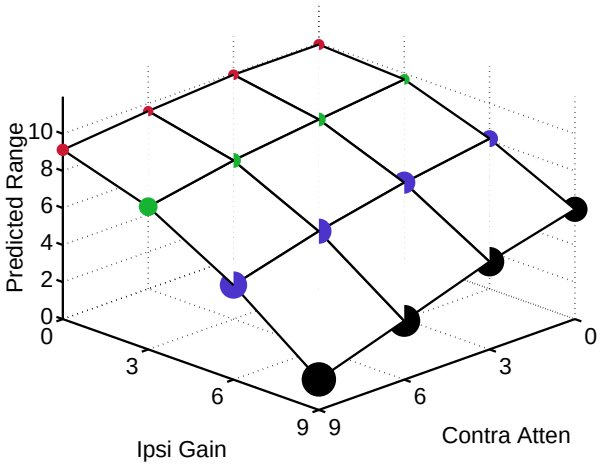

Fig. 4: Grand mean range ratings for one subject plotted as a 3D mesh over a control surface formed by contralateral attenuation levels and ipsilateral gain level. The size of the symbols correspond to the grand mean rating of spatial extent for the auditory image associated with each stimulus.

It is a straightforward application of Multiple Regression Analysis (MRA) to characterize this interaction in terms of a prediction equation. The inversion of the prediction equation allows a contralateral attenuation value to be determined so that a source at a desired loudness (which specifies the ipsilateral gain value) can also be positioned at a particular range, at least for whispered speech sources arriving at an incidence angle of 130° azimuth.[5]

4. Discussion

A general principle that helps to understand the observed range ratings is that one perception may sometimes depend upon another prior perception. For example, Gogel [35] made this same point about how visual cues to exocentric distance (i.e., differential displacements) depend upon the prior human judgments of perceived egocentric (absolute) distance:

> "[Because] differential displacements are indeterminate with respect to an exocentric distance without specifying an egocentric distance, a *perception* of egocentric distance is required to translate these cues to a *perception* of exocentric distance." ([35], p. 367)

It is also true that judgments of auditory range depend not only on perceptual factors, but also upon cognitive factors that influence human perceptual judgment. It is clear that response biases (tendencies) contribute to both absolute (egocentric) and relative (exocentric) distance judgments, and that listeners rely upon these biases more and more as the adequacy of the stimulus cues is reduced. In the absence of strong absolute cues to auditory range, all stimuli will tend to be perceived at some specific range, typically rather near the listener (Gogel [36] termed this tendency the *specific distance tendency*). Anechoic sound sources presented near the listener's median plane, produced at a variety of ranges greater than 1 m typically tend to be perceived at a range somewhat less than 1 m, even when the *proximal* SPL is quite low.

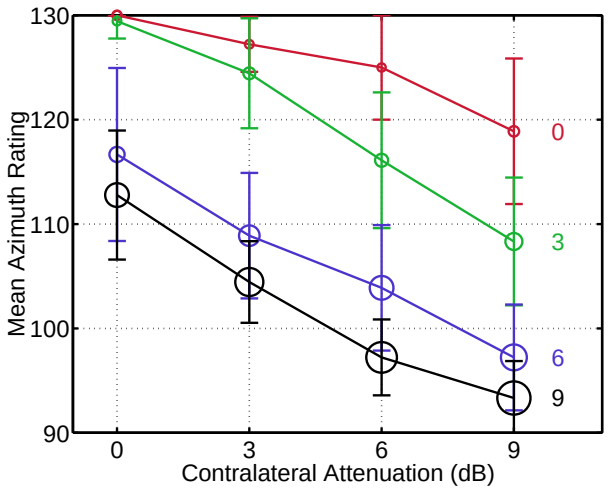

Fig. 5: Grand mean azimuth ratings for one subject plotted over contralateral attenuation levels, combining results for the same stimuli employed in the range rating experiment. The parameter of the graph again is ipsilateral gain level, the values of which are used to label each of the four sets of connected plotting symbols (numerals appearing just to the right of each). The size of the symbols again correspond to the grand mean rating of spatial extent for the auditory image associated with each stimulus.

Another perceptual attribute that likely depends upon both the manipulated stimulus parameters, and also the prior perception of other attributes (such as loudness), is the apparent azimuth angle of the virtual sound sources displayed using the technology under

[5] The details of the inverse prediction via MRA are beyond the scope of this short paper. Regression equation inversion in the multiple predictor case is well described elsewhere [34]. Suffice it to say that a multivariate linear equation can be inverted to provide required values of one predictor variable when the values of both the criterion variable and the other predictor variable are specified.

test. As large level differences are introduced between ipsilateral and contralateral signals (i.e., manipulating ILD), the apparent azimuth of the virtual source should also grow larger. Fig. 5 shows that this is indeed the case. Nevertheless, the covariation in perceived range and azimuth is not particularly a problem for the display, especially when the changes in apparent spatial extent of the sound source is considered. As all three of these perceptual attributes were rated for all stimuli presented in the current investigation, it is easy to explain the consequences. As a virtual source approaches very near to a listener's ear, it also seems to fill more of space. The whispered speech sources do not seem as if they were emitted from a point in space, but rather from an emitter that takes up some space (such as the mouthg of a human talker subtending a significant angle in nearby space). Judgments of the azimuth angle of the center of that extended source naturally regress toward the 90^o pole toward the side of the listener. Since the time difference between ipsilateral and contralateral signal arrival is on the order of natural ITD values, the extreme lateralization does not sound unnatural. In fact, the head-related cues employed by this display re-introduce a perceived naturalness that is typically missing in most head-phone based spatial auditory displays. The basis for this naturalness is that the auditory sensation of nearby sources are in fact referenced to the body of the listener in a way that is not provided by this "proximity" effect. The listener hears the presence of their own body in the displayed virtual source, and thus a hightened sence of presence can be experienced. This point is underscored by the following quotation taken from a 1993 Micritimes interview with Brenda Laurel on "Rethinking the Human-Computer Relationship:"

> "From the beginning we've defined computers as extentions of our minds, and we've tried to model them after our minds. But the model of mind that is reified in the computer is a disembodied mind ... The reason VR is important is because it says mind doesn't exist without body; the senses are part of the mind and we have to take our bodies with us." ([37], p. 79)

5. Conclusion

The manipulations in the above-described experiments clearly produce relatively strong cues to auditory range, but also demonstrate an interaction between two predictors of range, quantified as ipsilateral gain and contralateral attenuation. The results further showed that other perceptual attributes, particularly apparent azimuth angle and apparent spatial extent of the virtual sound source, are also influenced by the manipulation of these two parameters. Testing at other sound source azimuth angles is not yet complete, but even though the utility of the current inverse prediction may seem limited, the derived control structure providing independent variation of loudness and range may in fact match well the requirements of certain applications. For example, an application might require the delivery of a voice message from an angle outside of the user's visual field, and at a range close enough to cause the user to notice it immediately. Furthermore, the importance of the spoken message so presented can be indicated by how close to the user's ear the source seems to be located. It was also shown that the sound source can be placed near the ear without making it significantly louder, rendering it less likely to mask other simultaneous messages. It is concluded that the spatial auditory display technology tested here can provide effective control over the range of a virtual sound source, and that this enables a useful telecommunication feature termed the "dark voice channel." An additional advantage of the tested auditory display technology is that it provides a potential means for range-based speech segregation, as reported in [38]. Such display technology can provide a great improvement in close-range auditory imagery either with or without active head tracking, but its greatest impact will be on non-interactive systems that otherwise perform poorly due their failure to support active localization.

References

1. W. L. Martens and A. Yoshida, "Augmenting spoken telecommunication via spatial audio transformation," *The Journal of Three Dimensional Images*, vol. 14, no. 4, pp. 169–175, 2000.

2. W. L. Martens, "Display of nearby virtual sound sources outside the user's field of view: The dark voice channel," in *Proc. HC2001: 4^{th} Int. Conf. on Human and Computer*, Aizu-Wakamatsu, Japan, Sept. 2001, 3D Forum, pp. 207–215.

3. M. B. Gardner, "Distance estimation of 0^o or apparent 0^o-oriented speech signals in anechoic space," **J. A***cous.* **S***oc.* **A***mer.*, vol. 45, no. 1, pp. 47–53, 1969.

4. Frederic L. Wightman and Doris J. Kistler, "Resolution of front back ambiguity in spatial hearing by listener and source movement," **J. A***cous.* **S***oc.* **A***mer.*, vol. 105, pp. 2841–2853, 1999.

5. D. L. Budenz, *Atlas of Visual Fields*, Lippincott Williams & Wilkins, Publishers, 1997.

6. J. P. Szlyk, "Peripheral visual field loss and driving performance," Tech. Rep., AAA Foundation for Traffic Safety, 1992.

7. W. L. Martens, B. McRuer, C. T. Childs, and E. Virree, "Physiological approach to optimal stereographic game programming: A technical guide," in *Proc. IS & T/SPIE*, San Jose, CA, 1996, vol. 2653, Stereoscopic Displays and Virtual Reality Systems III.

8. USAF, *Vision in Military Aviation*, Wright Patterson Air Force Base, Fairborn, OH, USA, 1958.

9. S. R. Ellis, "Origins and elements of virtual environments," in *Virtual Environments and Advanced Interface Design*, W. Barfield and T. A. Furness III, Eds., chapter 2, pp. 14–57. Oxford University Press, 1995, ISBN 0-19-507555-2.

10. D. R. Begault, *3-D Sound for Virtual Reality and Multimedia*, Academic Press, 1994, ISBN 0-12-084735-3.

11. H. Moller, "Fundamentals of binaural technology," *Applied Acoustics*, vol. 36, pp. 171–218, 1992.

12. D. R. Begault and E. M. Wenzel, "Headphone localization of speech," *Human Factors*, vol. 35, no. 2, pp. 361–376, 1993.

13. J. F. Burger, "Front-back discrimintation of the hearing system," *Acustica*, vol. 8, pp. 301–302, 1958.

14. J.M̃. Loomis, C. Hebert, and J.G̃. Cicinelli, "Active localization of virtual sounds," **J. A***cous.* **S***oc.* **A***mer.*, vol. 88, no. 4, pp. 1757–1763, Oct. 1990.

15. D. S. Brungart and W. M. Rabinowitz, "Auditory localization of nearby sources: Head-related transfer functions," **J. A***cous.* **S***oc.* **A***mer.*, vol. 106, no. 3, pp. 1465–1479, 1999.

16. W. L. Martens and A. Yoshida, "Psychoacoustically – based control of auditory range: Display of virtual sound sources in the listener's personal space," in *Int. Conf. on Information Society in the 21st Century: Emerging Technologies and New Challenges (IS2000)*, Nov. 2000, pp. 288–294.

17. F. E. Toole, "In-head localization of acoustic images," **J. A***cous.* **S***oc.* **A***mer.*, vol. 48, pp. 943–949, 1969.

18. R. Sommer, *Personal Space – The Behavioral Basis of Design*, Englewood Cliffs, N. J.: Prentice-Hall Inc., 1969.

19. A. Yoshida and W. L. Martens, "Whisper function: An audio transformation for conveying a confided speech message in a multi-user virtual environment," in *IEICE Tohoku Branch Conf.*, University of Aizu, Aug. 2000.

20. N. I. Durlach, A. Rigpulos, X. D. Pang, W. S. Woods, A. Kulkarni, H. S. Colburn, and E. M. Wenzel, "On the externalization of auditory images," *Presence*, vol. 1, no. 2, pp. 251–257, Spring 1992, ISSN 1054–7460.

21. W. M. Hartmann and A. T. Wittenberg, "On the externalization of sound images," **J. A***cous.* **S***oc.* **A***mer.*, vol. 99, pp. 3678–3688, 1996.

22. D. R. Begault, E. M. Wenzel, A. S. Lee, and M. R. Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," in *Proc. Audio Engineering Society 108th Int. Conv.*, 2000, Preprint 5134.

23. N. Sakamoto, T. Gotoh, and Y. Kimura, "On "out-of-head localization" in headphone listening," **J. A***udio* **E***ng.* **S***oc.*, vol. 24, pp. 710–715, 1976.

24. R. V. L. Hartley and T. C. Fry, "The binaural localization of pure tones," *Physics Review*, vol. 18, pp. 431–442, 1922.

25. D. S. Brungart, "Near-field auditory localization," in *Proc. 3rd Int. Conf. on Auditory Display*, Palo Alto, CA, 1996.

26. G. von Békésy, "Über die Entstehung der Entfernungsempfindung beim Hören [On the origin of the sensation of distance in hearing]," *Akustische Zeitschrift*, vol. 3, pp. 21–31, 1938, [Available in English in Békésy, G. von (1960). Experiments in hearing (E. G. Wever, Ed.) (pp. 301-313). New York: McGraw-Hill.].

27. W. L. Martens, "Psychophysical calibration for controlling the range of a virtual sound source: Multidimensional complexity in spatial auditory display," in *Proc. Int. Conf. on Auditory Display*, Espoo, Finland, 2001, ICAD, pp. 197–207, Keynote address.

28. D. J. M. Robinson and R. G. Greenfield, "A binaural simulation which renders out of head localization with low cost digital signal processing of head related transfer functions and pseudo reverberation," in *Proc. Audio Engineering Society 104th Int. Conv.*, 1998, Preprint 4723.

29. R. O. Duda and W. L. Martens, "Range dependence of the response of an ideal rigid sphere," **J. A***cous.* **S***oc.* **A***mer.*, vol. 105, no. 5, pp. 3048–3058, 1998.

30. H. Chen, "The incorporation of range in a DSP-based HRTF model," Technical report of the department of Electrical Engineering, San Jose State University, 1998.

31. D. S. Brungart, "A speech-based auditory distance display," in *Proc. Audio Engineering Society 109th Int. Conv.*, Los Angeles, 2000.

32. P. A. Keating and M. K. Huffman, "Vowel variation in Japanese," *Phonetica*, vol. 41, pp. 311–322, 1984.

33. W. L. Martens and N. Zacharov, "Multidimensional perceptual unfolding of spatially processed speech I: Deriving stimulus space using INDSCAL," in *Proc. 109th Conv. of the Audio Engineering Society*, Los Angeles, Sept. 2000, Preprint 5224.

34. N. R. Draper and H. Smith, *Applied Regression Analysis*, John Wiley & Sons, New York, 1981.

35. W. C. Gogel, "The organization of perceived space," in *Indirect Perception*, Cambridge, Massachusetts, 1997, pp. 361–386, MIT Press.

36. W. C. Gogel and J. D. Tietz, "Absolute motion parallax and the specific distance tendency," *Perception and Psychopysics*, vol. 13, pp. 284–292, 1973.

37. M. Robin, "Rethinking the human-computer relationship, an interview with author Brenda Laurel," *Microtimes*, pp. 71–79, May 1993.

38. D. S. Brungart, "Distance-based speech segregation in near-field virtual audio displays," in *Proc. Int. Conf. on Auditory Display*, Espoo, Finland, 2001, ICAD, pp. 169–174.