Proceedings of the 2003 International Conference on Auditory Display, Boston, MA, USA, July 6-9, 2003

OPTIMIZING THE SPATIAL CONFIGURATION OF A SEVEN-TALKER SPEECH DISPLAY

Douglas S. Brungart

AFRL/HECB
2610 Seventh St.
WPAFB, OH 45433
douglas.brungart@wpafb.af.mil

Brian D. Simpson

AFRL/HECB
2610 Seventh St.
WPAFB, OH 45433
brian.simpson@wpafb.af.mil

ABSTRACT

Although there is substantial evidence that performance in multitalker listening tasks can be improved by spatially separating the apparent locations of the competing talkers, very little effort has been made to determine the best locations and presentation levels for the talkers in a multichannel speech display. In this experiment, a call-sign based color and number identification task was used to evaluate the effectiveness of three different spatial configurations and two different level normalization schemes in a seven-channel binaural speech display. When only two spatially-adjacent channels of the seven-channel system were active, overall performance was substantially better with a geometrically-spaced spatial configuration (with far-field talkers at -90°, -30°, -10°, 0°, +10°, +30°, and +90° azimuth) or a hybrid near-far configuration (with far-field talkers at -90°, -30°, 0°, +30°, and +90° azimuth and near-field talkers at ±90°) than with a more conventional linearly-spaced configuration (with far-field talkers at -90°, -60°, -30°, 0°, +30°, +60°, and +90° azimuth). When all seven channels were active, performance was generally better with a "better-ear" normalization scheme that equalized the levels of the talkers in the more intense ear than with a default normalization scheme that equalized the levels of the talkers at the center of the head. The best overall performance in the seven-talker task occurred when the hybrid near-far spatial configuration was combined with the better-ear normalization scheme. This combination resulted in a 20% increase in the number of correct identifications relative to the baseline condition with linearly-spaced talker locations and no level normalization. Although this is a relatively modest improvement, it should be noted that it could be achieved at little or no cost simply by reconfiguring the HRTFs used in a multitalker speech display.

1. INTRODUCTION

Many important communications tasks require listeners to extract information from a target speech signal that is masked by one or more competing talkers. In real-world environments, listeners are able to take advantage of the binaural difference cues that occur when competing talkers are located at different positions relative to the listener's head. This so-called "cocktail party effect" allows listeners to perform much better when they are listening to multiple voices in real-world environments where the talkers are spatially separated than they do when they are listening with conventional communications systems where the speech signals are electronically mixed together into a single signal that is presented monaurally or diotically over headphones.

Previous research has shown that the efficiency of multitalker communications can be greatly improved by audio displays that use digital filters called head-related transfer functions (HRTFs) to reproduce the binaural cues that normally occur when competing talkers are spatially separated [1, 2]. To this point, however, very little effort has been made to systematically develop an optimal set of HRTF filters capable of maximizing the number of talkers a listener can simultaneously monitor while minimizing the amount of interference between the different competing talkers in the system. Most experiments that have examined the effects of spatial separation on multitalker speech perception have placed the competing talkers at roughly equally spaced intervals in azimuth in the listener's frontal plane [3, 4]. One experiment [1] spatially separated the speech signals in elevation as well as azimuth, with elevation decreasing from +60° to -60° as the source location moved from left (+90° azimuth) to right (-90° azimuth). Another experiment [2] used a horizontal-plane source placement algorithm that maximized the absolute differences in the sine values of the azimuth angles of the talkers. And, more recently, a new talker configuration has been proposed in which the target and masking talkers are located at different distances (12 cm and 1 m) at the same angle in azimuth (90°) [5]. While it is possible to make theoretical arguments in favor of each of these possible talker configurations, at this point no clear consensus had been reached on how to best choose the locations of the talkers in multichannel speech display systems.

Another important issue that has thus far received little attention is how the relative levels of the competing voices in a multitalker display should be selected in order to optimize listener performance. In real-world environments, the levels of the talkers are determined by their production levels and their relative distances from the listener. In multichannel speech displays, the relative levels of the talkers can be influenced by a number of factors that are beyond the control of the display designer, including the production levels of the talkers, the sensitivity of the microphones used to record their voices, and the user-determined volume control settings of the intercom system. It is, however, possible to control the relative levels of the different talkers in a system that uses automatic gain control to equalize the input levels of the voices. At this point it is not clear whether the performance of a multichannel speech display could be improved by systematically adjusting the relative levels of the talkers.

In this experiment, we examined the effects of three different spatial configurations on the performance of a seven-channel multitalker speech display: a standard configuration where the talkers were evenly spaced in azimuth, a near-far configuration where two of the talkers were located very near the head, and a geometric

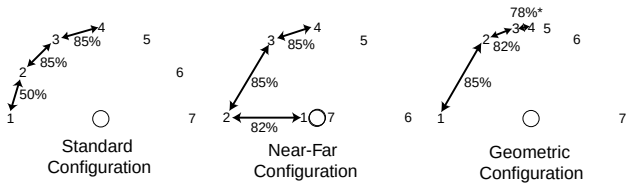

Figure 1: *Three spatial configurations for a system with seven competing talkers. The percentages on the arrows indicate performance in a two-talker CRM listening task with talkers located at the two endpoints of the arrows. See text for details.*

configuration where the talkers were concentrated near the median plane where listeners are known to be most sensitive to changes in sound source azimuth. Two different level normalization schemes were also evaluated in each of these configurations. The results are discussed in terms of their application to the design of improved multitalker displays.

2. METHODS

2.1. Speech Materials

The speech stimuli used in the experiment were drawn from the Coordinate Response Measure (CRM) corpus for multitalker research [6], which consists of sentences of the form "Ready, (Call Sign), go to (color) (number) now" spoken with all the possible combinations of eight call signs ("Baron," "Charlie," "Ringo," "Eagle," "Arrow," "Hopper," "Tiger," and "Laker"), four colors (red, blue, green, white), and eight numbers (1-8) by four male and four female talkers. In each trial of the experiment, the stimulus consisted of a combination of a target phrase, which was randomly selected from all of the phrases in the corpus with the call sign "Baron," and one or more masking phrases, which were randomly selected from the phrases in the corpus with call signs, colors, and numbers that differed from those used in the target phrase. These phrases were downsampled to 20 kHz from their original 40 kHz sampling rate, spatially processed by individually convolving them with the appropriate HRTF filters, mixed together electronically, and presented over headphones (Beyerdynamic DT-990) at a comfortable listening level (roughly 70 dB SPL). The task was to listen for the CRM phrase containing the target call sign "Baron", and then identify the color and number contained in that target phrase by using the mouse to select the color and number combination from a matrix of colored numbers on the CRT of the control computer, which was located in a quiet sound-treated listening room.

2.2. Spatial Configurations

Figure 1 illustrates the three seven-talker spatial configurations used in the experiment. The left panel of the figure shows the distribution of sources in the *standard* configuration, where the talkers were spaced every 30° in azimuth across the frontal hemisphere at a distance of 1 m. Similar source distributions have been used in previous studies where the talkers were distributed across seven talker locations but only 3-4 of the talkers were active at the same time [3, 4]. The middle panel of the figure shows the distribution of sources in the *near-far* configuration, with five "far-field" talkers geometrically spaced at -90°, -30°, 0°, +30°, and +90° azimuth at 1 m and two "near-field" talkers at plus and minus 90° in azimuth 12 cm from the center of the head. The right panel of the figure shows the distribution of sources in the *geometric* configuration where the sources were located at -90°, -30°, -10°, 0°, +10°, +30°, and +90° azimuth and a distance of 1 m.

The digital filters used to implement these three spatial configurations were derived from a set of HRTF measurements made on a KEMAR manikin with an acoustic point source [7]. These HRTFs were corrected for the response of the headphones used in the experiment (Beyerdynamic DT-990 measured with a KEMAR manikin) and used to generate 18-point linear-phase FIR filters at a 20-kHz sampling rate with the MATLAB FIR2 command. Then these filters were upsampled to 1 MHz, zero-padded in one ear to introduce the appropriate interaural time delay, and downsampled back to a 20 kHz-sampling rate. The resulting HRTF filters were convolved directly with the stimuli from the CRM corpus to generate the different spatial configurations used in the experiment.

2.3. Level Normalization

The two different level normalization schemes used in the experiment are illustrated in Figure 2. The white and black bars in the figure show the RMS levels in the left and right ears for speech-shaped noise signals that were spatially processed with the HRTFs for each of the seven source locations in each of the three spatial configurations in the experiment. The left column of the figure shows the levels in the default *center-of-the-head* normalization scheme, which adjusted the levels of the talkers so they would all be equally intense in the free field at the location of the center of the listener's head (with the head removed). This normalization had no effect on the relative levels of talkers at locations that were equidistant from the listener, but it did eliminate the roughly 18 dB increase that would have occurred at the 12 cm source locations of the near-far source configuration due to the decreased distance of the nearby talkers. In this normalization scheme, the levels in each ear varied naturally with talker location, and the talkers at the central locations (2-6) were always less intense than the most intense talker in either of the two ears.

The right column of Figure 2 shows the *better-ear* normalization scheme, which adjusted the overall levels of the talkers so that each one produced the same output level in the more intense ear. Thus, talkers located in the right hemisphere were all adjusted to have the same output level in the right ear, and talkers located in the left hemisphere were all adjusted to have the same output level in the left ear. In this normalization scheme, each talker was always as intense as the most intense talker in at least one of the two ears.

3. EXPERIMENT 1: INTERFERENCE BETWEEN ADJACENT SPATIAL CHANNELS

3.1. Methods

Experiment 1 was conducted as a subset of a more general experiment that examined the amount of interference between two competing CRM speech signals as a function of the spatial separation between the two stimulus locations. Within each block of trials, the first talker location was fixed at one of five angles (5°, 15°, 30°, 45°, 60° or 90°) at a distance of 0.12 m, 0.25 m, or 1 m, and the second talker location was varied from 0° to 90° at a

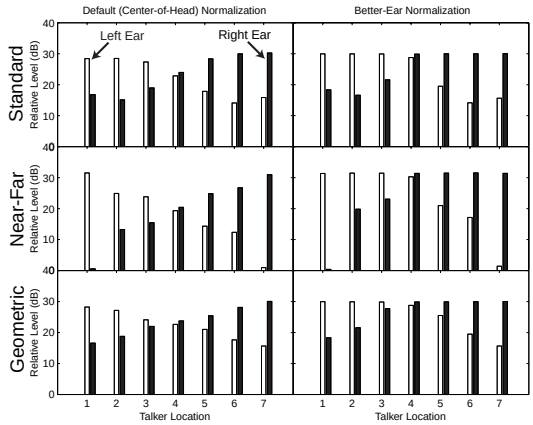

Figure 2: *RMS levels of spatially-processed speech-shaped noise in the left and right ears with the default and better-ear normalization schemes used in the experiment.*

distance of 1 m. In all cases, the two CRM phrases were spoken by the same talker, and a variation of the better-ear normalization scheme was used to adjust the levels of the two speech signals to have the same RMS level in the ear where the target speech was most intense[1]. Within each trial, the target phrase containing the call sign "Baron" was equally likely to originate from either of the two talker locations. Seven normal-hearing volunteer subjects served as listeners in the experiment (three male and four female), and each participated a minimum of 40 trials in each of the spatial configurations tested in the experiment.

3.2. Results

The double-headed arrows in Figure 1 show the percentages of correct color and number identifications in the conditions where the two competing talkers were located at adjacent source locations in one of the three spatial configurations tested in this experiment. For example, the 85% label on the arrow between source positions 3 and 4 in the standard configuration indicates that the listeners in Experiment 1 correctly identified both the color and the number in the target phrase in 85% of the trials where the two talkers were located at 0° azimuth and 30° azimuth and a distance of 1 m. The only exception is that the 78% value on the arrow between locations three and four in the geometric configuration (marked by an asterisk in the figure) represents performance for sources located at 5° and 15° in azimuth, and not the actual locations of 3 and 4 (0° and 10°) which were not directly tested in the experiment.

3.3. Discussion

The results of Experiment 1 highlight one of the major weaknesses of the standard linearly-spaced spatial configuration, namely that it fails to take into account the reduced sensitivity to changes in azimuth that occurs when sound sources are located near 90°. Mills, for example, showed that listeners are 6-10 times more sensitive to changes in the azimuth of sound sources near 0° than they are to changes in the azimuth locations of sounds near +-90° [8]. As a result of this reduced spatial sensitivity, the listeners had substantially more difficulty discriminating between talkers that were spatially separated by 30° near 90° azimuth (locations 1 and 2 in the left panel of Figure 1) than they did discriminating between talkers separated by only 10° near 0° azimuth (locations 3 and 4 in the right panel of Figure 1). Thus, on the basis of these results, one would expect listeners on average to respond correctly only 73% of the time if talkers happened to simultaneously occur on two randomly selected adjacent channels in the standard talker configuration, compared to 82% with the geometric configuration and 84% with the near-far configuration.

4. EXPERIMENT 2: PERFORMANCE WITH SEVEN SIMULTANEOUS TALKERS

4.1. Methods

The second experiment examined the effects of spatial configuration and level normalization the performance of a seven-channel speech display when all the competing talkers were active simultaneously. A total of seven different spatial configurations were tested in the experiment: all possible combinations of the three spatial configurations shown in Figure 1 and the two level normalization schemes shown in Figure 2, and a non-spatialized condition where all seven talkers were mixed together and presented to the listener diotically. Each block of 100 trials examined only one source configuration. Prior to each block, seven different talkers were randomly assigned to each of the locations in the selected source configuration. Once assigned, these talkers remained fixed at these source locations for the remainder of the block. Four of the talkers were male talkers from the CRM corpus, and the other three were female talkers from the corpus that were electronically processed to make their voices sound like natural male speech[2]. On the first trial of each block, one of the seven talkers was randomly selected to serve as the target talker. Then, after each subsequent trial of the experiment, there was a 25% chance that a different talker at a different location would be selected to serve as the target talker. In order to make the seven-talker CRM task less difficult, a 100 ms delay was introduced between the onset of the target phrase and the onset of the six masking phrases. This allowed the target phrase to stand out against the maskers in what would otherwise have been a nearly impossible task. It also reflected the fact that voices in real-world multitalker listening tasks are rarely, if ever, perfectly synchronized. A total of ten normal-hearing subjects served as listeners in the study, with each participating in 3-4 blocks of 100 trials in each of the seven conditions. Thus a total of 27800 trials were collected in the experiment.

4.2. Results

Figure 3 shows the percentages of correct responses for each of the seven target talker locations associated with each condition of the experiment. The three rows show the three different spatial configurations, and the two columns show the two different normalization schemes. Performance in the non-spatialized condition

[1] Note that this differs slightly from the better-ear normalization scheme described in section 2.3, in that it normalizes the level in the ear where the *target* talker is more intense and not necessarily the ear where each individual HRTF is more intense. This was done to eliminate the signal-to-noise advantage that normally occurs in one of the two ears when two talkers are spatially separated and focus exclusively on the binaural advantages of spatially separating two talkers.

[2] This processing was accomplished by using PSOLA synthesis to scale the F0 of the voices by a factor of 0.59 and the vocal tract sizes of the voices by a factor of 1.16.

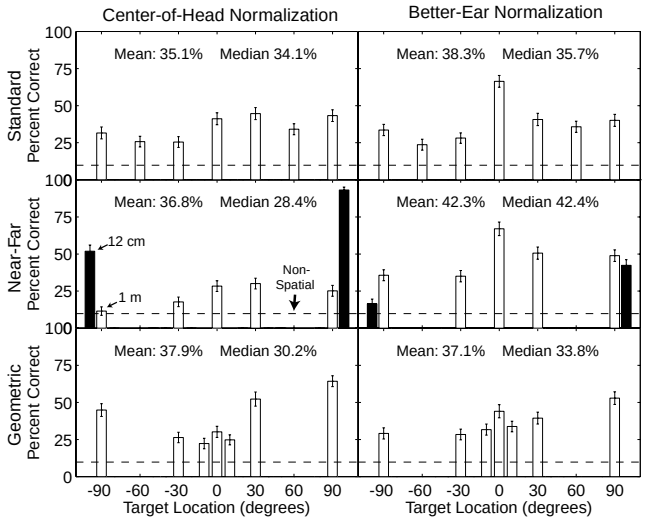

Figure 3: *Effects of spatial configuration and level normalization on performance in a seven-talker CRM listening task. Each panel also shows the mean and median percent correct across the seven locations in that condition. Differences larger than 1.1% in the overall mean values are statistically significant at the $p<0.05$ level. The error bars show the 95% confidence intervals of each data point.*

(9.8% correct) is indicated by the horizontal dashed line in each panel. The numbers at the top of each panel provide the overall mean performance in each condition as well as the median performance across the seven different talker locations in that condition.

When the baseline normalization scheme was used (left column of the figure), performance was roughly evenly distributed across the talker locations in the standard spatial location, but correct responses tended to be concentrated at the 12-cm locations in the near-far configuration and at the lateral locations ($\pm 90^{\circ}$) in the geometric configuration. The main effect of better-ear normalization (right column) was to reduce performance at the lateral locations where the talkers were attenuated and increase performance at the medial locations near 0° where the talkers were amplified (see Figure 2). Better-ear normalization also increased the overall number of correct identifications by 9% in the standard configuration and by 15% in the near-far configuration. Note also that all of the configurations produced better performance for talker locations in the right hemisphere (44% correct responses overall) than for talker locations in the left hemisphere (28% correct responses), which suggests the existence of a hemispherical asymmetry in the processing of spatially-separated speech channels.

5. DISCUSSION AND CONCLUSIONS

All six of the spatial configurations tested in Experiment 2 improved performance by more than a factor of three over the non-spatialized baseline condition. However, despite substantial differences in the locations and relative levels of the competing talkers in these six configurations, their overall mean performance levels varied over a relatively small range (35-42%). On one level, this suggests that listeners may not be particularly sensitive to the specific locations of the individual talkers in multitalker listening tasks with more than two talkers. Further research is needed to explore why this might be true. On a more practical level, however, it would be inappropriate to ignore the modest performance advantage that the near-far configuration with better-ear normalization had over the other configurations tested. That configuration produced mean performance that was 10% better than the standard configuration with better-ear normalization and more than 20% better than the standard configuration with center-of-head normalization. It also produced the highest median level of performance, indicating a reasonably even distribution across the seven talker locations (although it should be noted that performance was relatively poor for the 12 cm talker at -90°). When one considers that this performance improvement could be achieved with little or no cost simply by modifying the HRTFs used for the spatial processing, this configuration appears to warrant serious consideration by the designers of multichannel speech displays. Additional research is now needed to determine how this seven-talker configuration could be further improved, how it could be extended to configurations with fewer or more than seven talkers, and how other factors such as interactive head tracking might influence the best spatial configurations for multitalker speech displays.

6. ACKNOWLEDGEMENTS

This work was sponsored in part by AFOSR LRIR 01-HE-01-COR.

7. REFERENCES

[1] K. Crispien and T. Ehrenberg, "Evaluation of the 'cocktail party effect' for multiple speech stimuli within a spatial audio display," *J. of the Aud. Eng. Soc.*, vol. 43, pp. 932–940, 1995.

[2] W. T. Nelson, R. S. Bolia, M. A. Ericson, and R. L. McKinley, "Spatial audio displays for speech communication. a comparison of free-field and virtual sources," *Proc. of Hum. Fact. Erg. Soc.*, pp. 1202–1205, 1999.

[3] W.A. Yost, R.D. Dye, and S. Sheft, "A simulated 'cocktail party' with up to three sources," *Perc. Psychophy.*, vol. 58, pp. 1026–1036, 1996.

[4] M.L. Hawley, R.Y. Litovsky, and H.S. Colburn, "Speech intelligibility and localization in a multi-source environment," *J. Acoust. Soc. Am.*, vol. 105, pp. 3436–3448, 1999.

[5] D.S. Brungart and B.D. Simpson, "The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal," *J. Acoust. Soc. Am.*, vol. 112, pp. 664–676, 2002.

[6] R.S. Bolia, W.T. Nelson, M.A. Ericson, and B.D. Simpson, "A speech corpus for multitalker communications research," *J. Acoust. Soc. Am.*, vol. 107, pp. 1065–1066, 2000.

[7] D.S. Brungart and W.M. Rabinowitz, "Auditory localization of nearby sources. i: Head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 106, pp. 1465–1479, 1999.

[8] A.W. Mills, "On the minimum audible angle," *J. Acoust. Soc. Am.*, vol. 30, pp. 237–246, 1958.