Audio Engineering Society

Convention e-Brief

Presented at the 139th Convention
2015 October 29–November 1 New York, USA

Speech Intelligibility Advantages using an Acoustic Beamformer Display

Durand R. Begault[1], Kaushik Sunder[1,2], Martine Godfroy[1,2] and Peter Otto[3]

[1] NASA Ames Research Center, Moffett Field, CA 94035

Durand.R.Begault@NASA.gov

[2] San Jose State University Foundation

[3] Cal IT2, University of California San Diego

ABSTRACT

A speech intelligibility test conforming to the Modified Rhyme Test of ANSI S3.2 "Method for Measuring the Intelligibility of Speech Over Communication Systems" was conducted using a prototype 12-channel acoustic beamformer system. The target speech material (signal) was identified against speech babble (noise), with calculated signal-noise ratios of 0, 5 and 10 dB. The signal was delivered at a fixed beam orientation of 135° (re 90° as the frontal direction of the array) and the noise at 135° (co-located) and 0° (separated). A significant improvement in intelligibility from 57% to 73% was found for spatial separation for the same signal-noise ratio (0 dB). Significant effects for improved intelligibility due to spatial separation were also found for higher signal-noise ratios (5 and 10 dB).

1. BACKGROUND

The intelligibility advantage gained from spatial separation of multiple speech channels or "streams" is explained by the well-known "cocktail party effect" [1] and has been implemented in spatial auditory displays for teleconferencing or for radio communications for some time (e.g. [2]). The current study focuses on the use of a 12-channel phased array "beamformer" system to improve the intelligibility of a target communication stream heard simultaneously with an undesired communication stream. By strategically manipulating the positions of incoming communication streams in an auditory display, an intelligibility advantage should be gained.

2. SUBJECTS

Ten volunteer participants (ages 18-30) were recruited from the NASA Ames Human Systems Integration Division. Experimental procedures were compliant with Human Research Institutional Review Board protocols. All had normal hearing based on a screening questionnaire.

3. STIMULI GENERATION

Participants were seated at a distance of 1 m from a 12-channel acoustic beamform display comprised of a horizontal array of isolated 1 in drivers (modified Innovox MLA-ST-12 loudspeaker and custom amplifier hardware, "SoundBender" software from UC San Diego Cal-IT2, written within Max 7 software from Cycling '74). The height of the array was adjusted to ear level. The speech playback level was set to a level of 62 dB. Figure 1 shows the level of pink noise as a function of beam azimuth, measured at 1 m in front (90°). Figure 2 shows a diagram indicating the two beam azimuth directions (0° and 135°) used in the experiment. The experiment used a 15 s LAeq of speech material to calibrate levels relative to a 62 dB(A) playback level at the 135° position where the subject was seated.

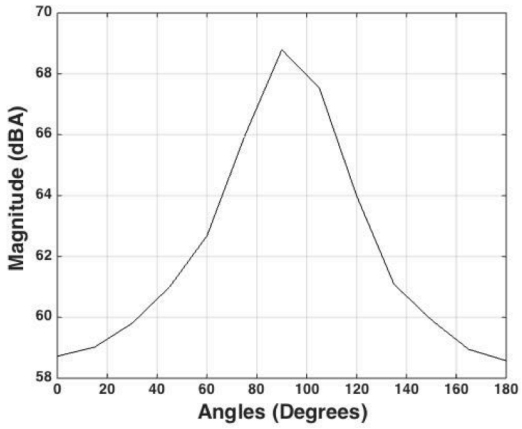

Figure 1 Level of Pink noise as a function of beam azimuth, measured at 1 m in front (90°).

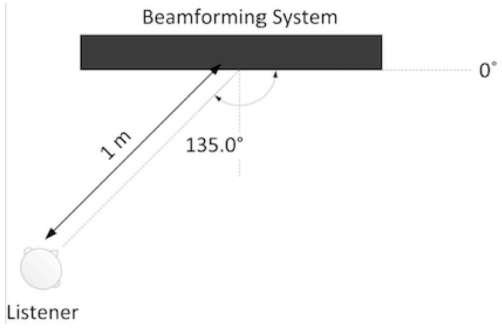

Figure 2 Configuration of listener, beamformer and stimuli angles used in the experiment.

4. EXPERIMENT TASK, CONDITIONS

The methodology used conformed to the Modified Rhyme Test (MRT) from ANSI S3.2 "Method for Measuring the Intelligibility of Speech Over Communication Systems" [3]. The test is a six-alternative forced choice of target words that differ in either their initial or final consonant. Within each of five experimental blocks, 50 trials of the six words corresponding to the 50 word sets of the MRT were presented. Blocks were repeated twice in random order.

The stimuli presented during each trial consisted of two simultaneous speech signals, as follows. A speech signal stimulus, i.e. a target word from the MRT list, duration ~1.0–1.2 s, was played at a random start time following the start of the block or after the completion of a trial. In addition, a multilayered stream of speech babble was played continuously during each block. The speech babble was randomly mixed from three male and three female talker recordings. Three of the talkers' words were uttered in reverse.

For each MRT word, six different talkers (three male, three female) were randomized between trials. Signals were always presented in the same direction towards the listener (135°) while the position of the babble was varied between 135° and 0° (ref. Figure 2).

Subjects indicated the target word that was heard from the six alternatives, which were presented on a touch screen surface (Apple iPad) immediately after the presentation of the target audio stimulus. Each answer was recorded by the experimental software, which then triggered the next stimulus.

A total of ten experimental blocks were presented to each subject in randomized order, to test spatial location and playback level as independent variables, in a repeated measures within-subjects design. Table I shows the spatial locations of signal and noise and their relative levels; a "block ID #" is provided as a shorthand reference to the condition. Each of these five blocks were evaluated twice by each subject. Each test resulted collectively in a total of 1000 responses per block type. Subjects took an average of 7-8 minutes to complete a block.

Block type #	Signal position	Noise position	Noise level	S/N ratio
1	135	135	62	0
2	135	0	52	10
3	135	0	62	0
4	135	135	57	5
5	135	0	57	5

Table I. Experimental conditions. The playback level of the signal was held constant at 62 dB(A).

Selected pairs of blocks were chosen to evaluate if there was a significant improvement in speech intelligibility as a function of spatial location and signal-noise (S/N) ratio. There are two spatially co-located condition blocks with the sound beam projecting both signal and noise to 135° and three non co-located conditions with the signal beam at 135° and noise at 0° (ref. Figure 2).

-Block 1 compared to block 2: the advantage, if any, of spatial separation of the speech babble to a different location. The spatial separation causes a net S/N increase due to a decrease in level of the babble by 10 decibels at the listener position (ref. Figure 1).

-Block 1 compared to block 3: if any, of spatial separation of the speech babble to a different location, with S/N adjusted to 0 decibels (by increasing the level of the speech babble by 10 dB). Unlike block 2, the S/N ratio in block 3 was the same as for block 1.

-Block 4 compared to block 5; the advantage, if any, of spatial separation of the speech babble to a different location, with the S/N ratio held constant at 5 dB.

We also checked the significance of S/N ratio with spatial location fixed (blocks 1 & 4, co-located spatial position; blocks 2, 3 and 5, non co-located spatial positions). Finally, we investigated differences in fixed S/N ratio and varied spatial position, and fixed spatial position and varied S/N ratio.

Subjects also rated their confidence in their responses after completing each block. A ten-point scale was used, anchored by the words "minimally confident" (0) and "maximally confident" (10). This allowed evaluation of the relationship between objective and subjective performance.

5. RESULTS

5.1. Effect of Spatial Separation

With the S/N ratio of 10 dB (comparison of block 1 to block 2), spatial separation of the signal from speech babble resulted in an increased rate of correct responses from 64.2% to 91.4%, a difference that was statistically significant ($X_1^2 = 214.55, p < .0001$).

With the same S/N ratio 0 dB (comparison of block 1 to block 3), spatial separation of the signal from speech babble resulted in a significant increase in the percentage of correct responses, i.e. 64.2% to 77.1% ($X_1^2 = 40.31, p < .0001$).

With the same S/N ratio of 5 dB (comparison of block 4 to block 5), spatial separation of the signal from speech babble resulted in a significant increase in the percentage of correct responses from 82.8% to 86.3% ($X_1^2 = 4.68, p = .03$).

5.2. Effect of S/N ratio

With spatial location fixed at 135° and noise at 135° (comparison of block 1 to block 4), the increase in S/N ratio from 0 to 5 dB led to significantly greater percentage of correct responses, 64.2% to 82.8% ($X_1^2 = 89.08, p < .0001$).

With spatial location of signal at 135° and noise at 0°, the increase in S/N ratio from 0 dB (block 3) to 5dB (block 5) to 10 dB (block 2) increased significantly the percentage of correct responses from 77.1% to 86.3% to 91.4% ($0, 5$ dB: $X_1^2 = 28.30,\ p < .0001$; $5, 10$: $X_1^2 = 13.12, p < .0001$).

Intelligibility is scored in terms of the percentage of adjusted correct answers out of the total number of responses. All data were corrected for the possibility of guessing per [3] by implementing the formula (adjusted correct answers = correct answers – (wrong answers / 5)). The corrected values are shown in Table II.

Block type #	Raw Score %	Adjusted Score %
1	64.2	56.9
2	91.4	89.6
3	77.1	72.5
4	82.8	79.3
5	86.3	83.5

Table II. Adjusted percentages for intelligibility by block type.

A Kruskal-Wallis H test was conducted to determine if there were significant differences in the confidence scores rated by the subjects.

The mean rankings increased 1824–2065–3068, respectively, for S/N ratios of 0, 5, and 10 dB, all significantly different (0, 5: $X_1^2 = 581.66,\ p < .0001$; 5,10: $X_1^2 = 18.14,\ p < .0001$).

Significant differences were also observed on comparing the mean confidence scores across the different spatial locations. Indeed, the mean ranking was significantly higher for spatially separated positions (0°: mean rank = 2779, 135°: mean rank = 2080, $X_1^2 = 309.99,\ p < .0001$).

We were interested in whether or not accuracy of responses was affected by the initial or final consonant sounds in the six alternative forced choice. The six words in each MRT ensemble differ only in the initial or final consonant. The results showed that the rate of correct responses was significantly higher by 6% when the initial consonant was different ("thaw" vs. "jaw") than when the final consonant was different ("sum" vs. "sun") (initial: 83.3%, final: 77.3%; $X_1^2 = 29.10,\ p < .0001$).

Further analyses of initial versus final consonant advantage showed no significant difference for S/N of 0 dB for the S135, N0 condition. At S/N 5 and 10 dB, there were significant differences ($p < .05$ or better) with MRT ensembles with the initial consonant difference having greater intelligibility. For all S135, N135 conditions, the S/N comparisons were significant favoring the initial consonant difference ($p < .05$ or better).

6. DISCUSSION

The increase in intelligibility as a function of S/N is not surprising. However, we found that with a fixed S/N ratio, spatial separation contributes a ~16% increase in intelligibility. This indicates an inherent advantage to a spatial auditory display using beamforming independent of gain adjustment. The spatial separation with its inherent S/N advantage represents an additive effect, yielding a ~33% intelligibility advantage.

Subjective ratings of confidence in intelligibility were found to parallel objective increases in intelligibility. We also found that intelligibility was increased when the initial consonant of the MRT word ensemble was different, compared to those ensembles where the final consonant was different. Similar results have been reported in the literature [4].The only exception to this was when the S/N ratio was 0 and the spatial locations were separated.

7. ACKNOWLEDGMENTS

This work was supported by NASA's System-wide Safety Assurance Technologies (SSAT) Project; and a NASA Space Act Agreement between CalIT^2 and UC San Diego. We appreciate the assistance of Mark R. Anderson and our colleagues in the Advanced Controls and Displays Laboratory at NASA Ames' Human Systems Integration Division.

8. REFERENCES

[1] E. C. Cherry, "Some experiments on the recognition of speech with one and two ears," *J. Acoustical Soc. Am.*, vol. 25, pp. 975-979.

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

[3] ANSI S3.2-2009, "Method For Measuring The Intelligibility Of Speech Over Communication Systems", Melville, N.Y.: Acoustical Soc. of America.

[4] M. A. Redford and R. L. Diehl, "The relative perceptual distinctiveness of initial and final consonants in CVC syllables" *J. Acoustical Soc. Am.*, vol. 106, pp. 1555-1565 (1999).