Proc. of the National Conference on Communications (NCC), Jan 2011, Bangalore, India

Azimuth-dependent Spatialization for a Teleconference Audio Display

N Harikrishnan Potty, Dipti Sengupta, Rajbabu Velmurugan, Preeti Rao
Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai, 400076, India
Email: {hkpottyn, rajbabu, prao} @ ee.iitb.ac.in, dipti.sengupta @ gmail.com

Abstract— **This paper proposes a novel scheme for the spatial rendition of audio in a teleconferencing situation. Several researchers have examined and established that there is a substantial improvement in intelligibility by the use of spatialization in such a multitalker environment, over monaural rendition. We provide experimentally obtained values for the Minimum Audible Angle (MAA) that provides release from masking in a speech over speech masking scenario. The results clearly indicate that the MAA varies with the azimuth. This dependency on azimuth has been utilized to improve the intelligibility of speech in a given teleconference within a limited auditory space. The advantage of such a scheme over simple auditory space division is demonstrated by another set of perceptual experiments.**

Keywords- minimum audible angle; MAA; spatial audio; teleconference; informational masking; energetic masking.

I. INTRODUCTION

Teleconferencing and telepresence systems are being increasingly used in social, academic, business, entertainment, military and many other areas. The objective is to design a spatial audio display to optimally render a teleconference system with small or large number of participants in a limited auditory space. The major advantages of such a system would be the improvement in intelligibility with a more natural and less stressful teleconferencing experience, using minimal hardware, namely a computer and a pair of headphones.

A typical teleconference is a complex multitalker environment, closely simulating the "cocktail party effect" [1]. The binaural advantage in the cocktail party effect has been investigated by several researchers. For example, advantages are provided by the spatial separation of speech from competing noise. Although these experiments have evaluated the advantages of spatialization with various kinds of noise maskers, the findings are not directly applicable to the teleconferencing scenario. This is because interfering speech signals and interfering noise signals produce different kinds of masking. Interfering noise signals produce only "energetic" masking, while interfering speech signals may produce both "energetic" and "informational" masking [2]. In this context, energetic masking refers to the concept of masking where the interfering signal overlaps in time and frequency with the target signal in such a way that portions of the target signal are rendered inaudible. Informational masking refers to the interference that occurs when the target and masker signals may or may not overlap in time and frequency but the listener is still unable to discriminate the contents of the target signal from the contents of a similar-sounding masker [2].

There can be several methods to provide release from both kinds of masking, i.e. to decrease the effect of the masker, and render the target audible and intelligible. A simple method is to apply a gain ratio, i.e. to change the target-to-masker ratio (energy or amplitude) to a value greater than unity. However, this is not practical when emulating a natural teleconference, where the targets and maskers can continuously change. The other method is to obtain spatial release from speech-speech masking. In case of multiple competing talkers, the advantage of spatial separation over monaural for release from masking has been reported by many [2][3].

It has been suggested in [3][4] that listeners are able to localize speech signals when two competing speech signals were spatially separated in azimuth. This localization was better compared to the case of a speech and noise signals which are spatially separated. The reason was that the listeners were able to use the differences in the apparent locations of the two sounds to reduce the informational component of speech-on-speech masking. A summary of various kinds of spatial locations used in multitalker displays is given in [5]. Most of these are for small number of participants and employ a simple equal division of the auditory space.

Brungart and Simpson [6] conducted experiments in which seven speakers were arranged in the horizontal plane separated by certain azimuths to improve speech display in the multitalker environment. Their results showed that there exists an advantage by spatially separating the speakers. They proposed a geometrically-spaced spatial (with speakers at -90°, -30°, -10°, 0°, 10°, 30°, 90° along the azimuth) configuration that improved the intelligibility when compared to a standard linearly-spaced (with speakers at -90°, -60°, -30°, 0°, 30°, 60°, 90° along the azimuth) configuration for speakers. However, the rationale behind the specific choice of angles for the geometrically-spaced spatial configuration was not provided, and hence the results are difficult to generalize to a different number of speakers. The present work is closely related to the work presented by [6]. We propose a speech display with a spatial configuration that exploits the dependence of minimum audible angle (MAA) on azimuth.

Our next objective was to find out the smallest angular or spatial separation that can provide the required release from masking in the case of two simultaneous speech signals. The earliest experiments to measure MAA were done in [7] with pure tones where it was observed in that the auditory resolution of space is not uniform in terms of the perceived azimuth. It has been observed that under ideal conditions, humans can localize single tone sounds or broadband noise pulses within a few degrees of accuracy and the value of this separation increases as we move away from the frontal plane [7][8][9]. We seek to extend the concept of MAA to speech, a broad bandwidth signal, thus adapting it for use in the teleconference scenario. In this context, we use the term MAA to refer to the minimum angular separation between two simultaneous speakers (one considered the target and the other, masker) so that a listener can correctly perceive the target speaker's speech based on some measurable criterion.

In the section II we present experiments conducted to study the variation of MAA along the azimuth for speech signals. Once the relation between azimuth and MAA is established, it is used to design an efficient spatial audio display system for a specified number of speakers in a given auditory space as discussed in Section III. The systems are evaluated in terms of achieved intelligibility in the presence of simultaneous speech. This is a measurable indicator that can be interpreted to represent the effectiveness and ease of the teleconferencing experience. In the concluding section, we summarise our observations and suggest possible future work.

II. Experiment 1: Minimum Audible Angle With Two Simultaneous Talkers

A. Stimuli

The audio data used for the perceptual experiments was a simulated version of the Coordinate Response Measure (CRM) database [10] and will be termed here on as the Simulated CRM (SCRM) database. It provides low contextual information and facilitates the variability necessary for repeated testing with the same listeners. The phrases are of the form, "Ready NAME, go to COLOUR NUMBER now". A set of eight common bisyllabic names were used, .e.g., 'Kanchan', 'Abha', 'Geeta', 'Rahul' etc., four common colours (Red, Blue, Green and Brown) and numbers from one to five and ten were used. The number 'six' was avoided after considering the listeners' feedback that 'six' was easily identifiable compared to other numbers. Thus a total of 8x4x6 = 192 combinations for each speaker were recorded. Five female speakers were used for recording thus yielding 192x5 = 960 phrases in total. The use of same gender speech for target and masker increases the difficulty level of the test by restraining the use of gender-specific cues by the listener for speech unmasking [2].

The utterances were spatialized by convolving them with HRTFs corresponding to various azimuth positions. The success of binaural synthesis and the quality of the generated spatial sound depend on the HRTF used. Obtaining these HRTFs from actual measurements and post-processing them for use in spatialization is a time-consuming process which requires specialized equipments and setup. Hence most binaural rendition systems use measured HRTFs provided by the MIT Media Lab [11]. These HRTFs are available for various azimuth positions and elevations. The azimuth positions were 5° apart at 0° elevation, the elevation that we are interested in. To better utilize the available auditory space we generated HRTFs for azimuths that are 1° apart. We made use of the interpolation technique mentioned in [12] to generate HRTF for those azimuths for which standard HRTFs are not available. In this approach, the available HRTFs are converted into their minimum-phase versions and the corresponding Interaural Time Delays (ITDs). These two components are then separately interpolated and combined back to get the required HRTF.

The audio phrases were sampled at 44.1 kHz and the average duration of an utterance was about 5.2 seconds. The phrases corresponding to two speakers were presented to the listeners. These phrases were word aligned at the start, i.e., at the word 'Ready' and also at the onset of the COLOUR and NUMBER. All speech utterances were Group Amplitude Normalized using Cool Edit Pro. 2.0 so as to ensure the same Target-to-Masker ratio of 0 dB (or TMR) throughout the experiment. The word alignment was done to avoid the precedence effect [13]. The normalization with respect to the amplitude was done so as to prevent the listener from taking advantage of the level difference cue [2].

B. Experiment Design

In this experiment, two audio files from two different speakers from the SCRM database were played simultaneously from two different spatial locations. As mentioned earlier, these phrases were spatialized by convolving them with HRTFs corresponding to various azimuth positions. One of the speakers was the target and the other masker. The target was randomly chosen during a test.

Audio was rendered over a pair of Sennheiser HD 201 circum-aural headphones. The listener was required to follow the target by the 'Name' cue. Given the name, the listener was required to follow the target speaker and note down the COLOUR and NUMBER spoken by the same. The file playing sequence was completely randomized with respect to the angular separations at each azimuth. All the testing has been done at azimuths on the left quadrant of the frontal plane. This has been done to obtain the worst case data, as listeners tend to perform better when the speakers are on the right quadrant [6]. The azimuths and the corresponding angular separations considered in the experiment are shown in Table I. The angular separations for each azimuth were chosen after pre-testing with a larger number of angles. For testing a given azimuth, each angular separation was repeated five times. For each azimuth and angular separation the positions of the target and the masker were interchanged and tested. Thus, for azimuth = 0°, there were total 6 angular separations hence 6*5*2 = 60 test files.

C. Results

The experiment was conducted with 10 normal-hearing adult listeners. Correct identification of the COLOUR or NUMBER spoken by the target was considered success. Accordingly we define a measure termed "separability" as

TABLE I. AZIMUTH CENTRES AND ANGULAR SEPARATIONS USED

Azimuth centre	Angular separations tested
0°	1°, 2°, 3°, 4°, 5°, 6°
-30°	4°, 6°, 8°, 10°, 12°, 14°
-45°	10°, 12°, 15°, 18°, 20°
-60°	15°, 20°, 22°, 25°

$$separability = \frac{C}{M} * 100\ \%, \quad (1)$$

where C is the number of correct identifications (in COLOUR or NUMBER, counted separately) and M the total number of instances. This M (200 in our experiments) is twice the number of times the target-masker pair of audio at a particular azimuth and angular separation is presented to all listeners. The multiplicity factor is to take care of both the COLOUR and NUMBER. The minimum angular separation at which the "separability", defined above, was greater than a threshold fixed percentage was noted. This minimum angular separation is denoted by MAA. Fig. 1 shows how the minimum angular separations vary with azimuth for different values of separability (50%, 60%, 70% and 80%). From Fig 1, it is clear that the minimum spatial separation between the target and the masker for the release of informational masking has an almost linear increase with azimuth. At a fixed azimuth, the achieved separability decreases as the angular separation decreases.

Fig. 1 is based on combined percentage correct across listeners. It is important also to understand the differences between listeners in terms of separability achieved for a given azimuth and angular separation. Fig. 2 depicts this in terms of the mean and standard deviation of the set of angular separations measured separately for each listener to achieve 80% separability for specified azimuth. We observe that although there is considerable overlap in the across-listener MAA ranges for different azimuths, the trend of increasing MAA with azimuth is clearly visible. This indicates that in spite of individual variability, it should be possible to exploit the characteristic dependence of MAA on azimuth for all listeners. Finally, we pick the average MAA values obtained at 80% separability in Fig. 1 for the proposed design of our teleconferencing system. These values are provided in Table II.

D. New Scheme for Placement of Speakers

From this experiment it can be concluded that the angular separation required for release from informational masking increases as we move away from the frontal plane. The experiment also gives some quantitative data regarding the amount of angular separation required at specific values of azimuths. This data can be utilized to place the speakers efficiently in a speech display. In the next section, we propose a novel scheme for placing participants in a spatial teleconference, based on the results of experiment 1, which can accommodate much larger number of participants in the limited auditory space still providing the spatial advantage of release from masking. We illustrate a general procedure for achieving this via an example.

III. EXPERIMENT 2: AZIMUTH DEPENDENT SOURCE PLACEMENT

A. Stimuli

The audio data used for this experiment is the same SCRM dataset used in the previous experiment.

B. Experiment Design

The experiment was conducted to show the advantage of an azimuth dependent speaker placement scheme over a scheme in which the speakers were spaced at equal azimuths. Generally the number of speakers that fit in a specific area depends on the total number of speakers. This becomes more

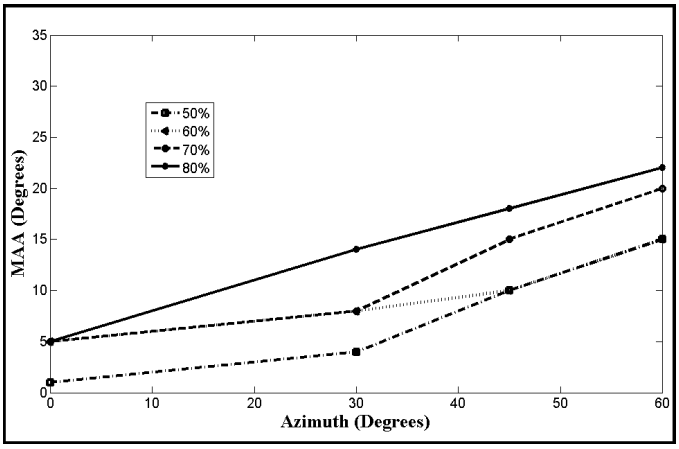


Figure 1. Variation of MAA with azimuth.

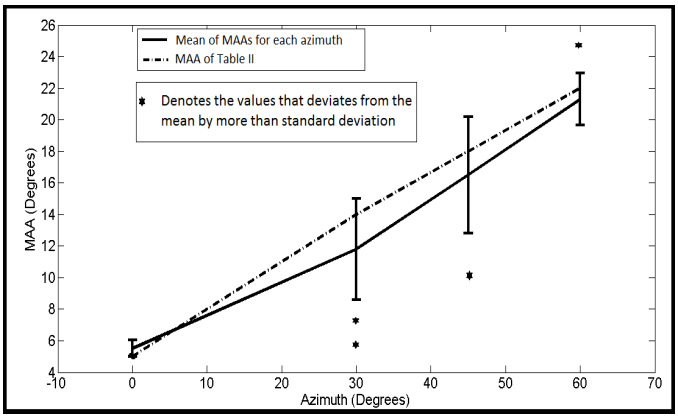


Figure 2. Average separation between the target and the masker required for the release of informational masking for each azimuth. The two bars (on top and bottom of the solid line) denote the standard deviation. The MAA of listeners for whom the deviations from the mean values, for the corresponding azimuth, are more than the standard deviation appear as outliers.

TABLE II. MAA FOR VARIOUS AZIMUTH CENTRES

Azimuth	MAA
0°	5°
-30°	14°
-45°	18°
-60°	22°

relevant as the projected locations of the speakers are not to change over the course of time. For our experiment, a large teleconference scenario is assumed where we have to fit in a certain number of speakers within certain spatial angle region.

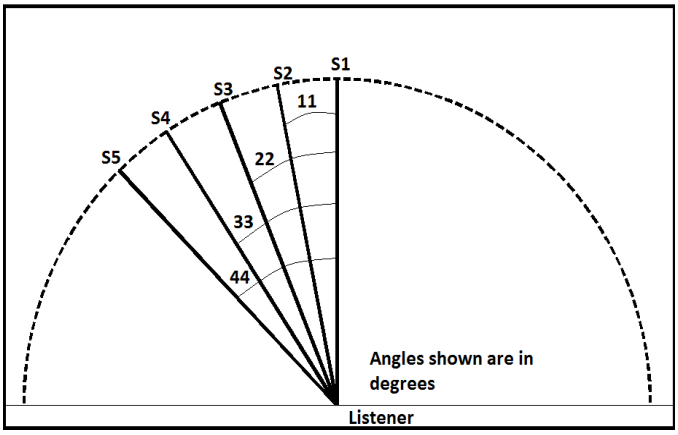
In this experiment we have considered the spatial region between 0° and -45° azimuths (Fig. 3). Assume that 5 speakers need to be fit in this region.

The experiment compares two schemes for speaker placement. In Scheme-1 of speaker placement (Fig. 3.a), the speakers are placed in a linear configuration with equal angular separation between the neighbours. In Scheme-2 (Fig. 3.b) the speakers are placed after taking into consideration the MAA and its variation along the azimuth as represented by Fig. 1. The speakers are placed at angular separations to span the available total angle and such that the separability (i.e. % correct) achieved is similar across all pairs of adjacent speakers. The azimuth values chosen for both the schemes are shown in Table III.

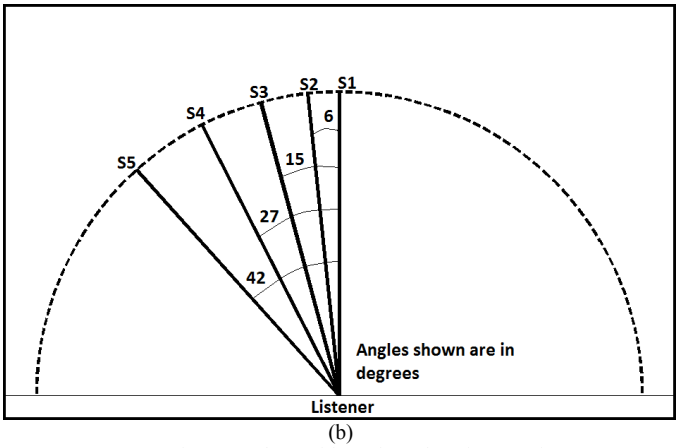
The speaker locations are selected based on the Fig. 1. As the number of speakers who need to be fit into any given area increases, the percentage separability that can be achieved decreases. In the 5 speaker case it is possible to achieve the 80% separability. Also best attempts have been made to select the positions such that the standard (non-interpolated) HRTFs can be used. The speakers were arranged as shown in Fig.3.b.

TABLE III. THE SPEAKER LOCATIONS ARE SELECTED BASED ON THE FIG. 1. AS THE NUMBER OF SPEAKERS WHO NEED TO BE FIT INTO ANY GIVEN AREA SPEAKER LOCATIONS FOR THE TWO SCHEMES

	Scheme-1:Linear and equally spaced speaker locations	**Scheme-2:Azimuth dependent speaker locations**
Speaker Locations	0°, -11°, -22°, -33°, -44°	0°, -6°, -15°, -27°, -42°



(a)



(b)

Figure 3. The position of speakers in Experiment 2. (a) Scheme 1 (b) Scheme 2

In this experiment, the audio files corresponding to two nearby speakers' locations from the SCRM database were played simultaneously at the chosen azimuths from Table III. Two neighbouring speakers were selected randomly for this purpose. One of the speakers was the target and the other masker. The speakers were all female as different gender target-masker case was found to be considerably easier. Audio was rendered over a pair of Sennheiser HD 201 circumaural headphones. As in the previous experiment, the listener was required to follow the target by the 'Name' cue. Given the 'Name', the listener was required to follow the target speaker and note the COLOUR and NUMBER spoken by the same.

C. Results

This second experiment was conducted with 8 listeners. The objective of the experiment was to compare intelligibility of speech obtained using the Scheme-1 and Scheme-2 for speaker placement. The intelligibility was measured using the separability percentage, which is the same as the one defined in (1). The separability obtained from Scheme-1 is shown in Table IV and that obtained from Scheme-2 is shown in Table V. The location of the speakers and the angle around which they are placed are mentioned in the first column in bold and normal fonts, respectively.

It is clear from the results that Scheme-2, which uses azimuth dependent placement of speakers, has an improved separability at most azimuths, when compared to the Scheme-1 where speakers are linearly and equally placed in azimuth. Further, the separability (measure of intelligibility) is more uniform across azimuth in Scheme II.

The results clearly show that the azimuth dependent scheme (Scheme-2) that places speakers considering the MAA can place more number of speakers within a limited auditory space, without significantly affecting the overall speech intelligibility. In a placement scheme that does not exploit this azimuth dependency of MAA (Scheme-1) the lowest performance tends to fall off rapidly as the number of speakers increase. The advantage of the azimuth-dependent scheme is expected to become even more significant when we move away from the frontal plane where the separation between the target and the masker should be higher.

TABLE IV. SCHEME-1:SEPARABILITY WITH LINEAR AND EQUALLY SEPARATED SPEAKER LOCATIONS (IN THE FIRST COLUMN, BOLD ONES INDICATE THE POSITION OF SPEAKERS AND THE NORMAL ONES THE AZIMUTH AROUND WHICH THE TESTING IS DONE).

Speaker locations	Separability(%)
0°,-5.5°,**-11°**	87
-11°,-16.5°,**-22°**	89
-22°, -27.5°,**-33°**	79
-33°,-38.5°,**-44°**	82

TABLE V. SCHEME-2:SEPARABILITY WITH AZIMUTH AND MAA DEPENDENT SPEAKER LOCATIONS (IN THE FIRST COLUMN, BOLD ONES INDICATE THE POSITION OF SPEAKERS AND THE NORMAL ONES THE AZIMUTH AROUND WHICH THE TESTING IS DONE).

Speaker locations	Separability(%)
0°,-3 °,**-6 °**	88
-6 °,-10.5 °,**-15 °**	89
-15 °,-21°,**-27 °**	91
-27 °,-34.5 °,**-42 °**	91

IV. CONCLUSIONS

In this work, we have experimentally shown the variation of minimum audible angle (MAA) with azimuth for the scenario where the target and masker are speech signals. This scenario is applicable to teleconferencing systems where audio is rendered over headphones. The experimentally measured variation was used as a basis for the design of an efficient system for placing participants in an audio teleconference. The proposed design method locates a given number of participants in a specified restricted auditory space so that speech intelligibility is least compromised. It is expected that the measured improvements in speech intelligibility will result in corresponding improvements in the effectiveness and ease of the overall teleconference experience. The proposed speaker placements can be further improved by considering the modification of relative signal levels. Also the variation of MAA with the elevation is not considered within the scope of this work. These aspects as well as reducing the computational complexity of spatial rendering will be part of future work.

ACKNOWLEDGMENT

The research work is supported by the project "National Program on Perception Engineering", sponsored by the Department of Information Technology, MCIT, Government of India.

REFERENCES

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and two ears", *J. Acoust. Soc. Am., 25*, (Sep. 1953), 975-979.

[2] D. S. Brungart, B. Simpson, M. Ericson, and K. Scott, "Informational and energetic masking effects in the perception of multiple simultaneous talkers", *J. Acoust. Soc. Am., 110 (5)*, Pt.1, (Nov. 2001), 2527-2538.

[3] R. Freyman, K. Helfer, D. McCall, and R. Clifton, "The role of perceived spatial separation in the unmasking of speech", *J. Acoust. Soc. Am*., 106 (6), (Dec. 1999), 3578–3587.

[4] H. S. Colburn, B. G. Shinn-Cunningham, G. Kidd, Jr., and N. Durlach, "The perceptual consequences of binaural hearing", International Journal of Audiology 2006; 45 (Supplement1): S34 - S44D.

[5] D. S. Brungart, and B. Simpson, "Improving Multitalker Speech Communication with Advanced Audio Displays", pp. 30-1 to 30-18, Meeting Proceedings RTO-MP-HFM-123, Paper 30, Neuilly-sur-Seine, France, 2005.

[6] D. S. Brungart, and B. Simpson, "Optimizing the spatial configuration of a seven-talker speech display", in Proceedings of the International Conference on Auditory Display (Boston, MA, USA, July6-9, 2003).

[7] A. W. Mills, "On the minimum audible angle", *J. Acoust. Soc. Am.,* 30 (4), (Apr. 1958), 237-246.

[8] S. Carlile, P. Leong, and S. Hyams, "The nature and distribution of errors in sound localization by human listeners", *Hear. Res.,* 114 (1-2), (Dec. 1997), 179–196.

[9] V . Best, S. Carlile, C. Jin, and A. van Schaik, "The role of high frequencies in speech localization", *J. Acoust. Soc. Am.*, 118 (1), (July 2005), 353–363.

[10] R. Bolia, W. Nelson, M. Ericson, and B. Simpson, ''A speech corpus for multitalker communications research'', *J. Acoust. Soc. Am*, 107, (Feb 2000), 1065–1066.

[11] http://sound.media.mit.edu/resources/KEMAR.html (Last accessed: 20/09/2010)

[12] L. Chen, H. Hu, Z. Wu , "Head-related impulse response interpolation in virtual sound system", Fourth International Conference on Natural Computation, 2008.

[13] R. Freyman, U. Balakrishnan, and K. Helfer, "Spatial release from informational masking in speech recognition", *J. Acoust. Soc. Am.,* 109 (5), Pt. 1, (May 2001), 2112-2122.