applied sciences

MDPI

Communication

Spatial Unmasking Effect on Speech Reception Threshold in the Median Plane

Nathan Berwick and Hyunkook Lee *

Applied Psychoacoustics Laboratory (APL), University of Huddersfield, Huddersfield HD1 3DH, UK; nathanjpberwick@icloud.com

* Correspondence: h.lee@hud.ac.uk; Tel.: +44-1484-471893

Received: 15 July 2020; Accepted: 26 July 2020; Published: 30 July 2020

Abstract: This study examined whether the spatial unmasking effect operates on speech reception thresholds (SRTs) in the median plane. SRTs were measured using an adaptive staircase procedure, with target speech sentences and speech-shaped noise maskers presented via loudspeakers at −30°, 0°, 30°, 60° and 90°. Results indicated a significant median plane spatial unmasking effect, with the largest SRT gain obtained for the −30° elevation of the masker. Head-related transfer function analysis suggests that the result is associated with the energy weighting of the ear-input signal of the masker at upper-mid frequencies relative to the maskee.

Keywords: spatial unmasking; speech reception threshold; median plane

1. Introduction

It is widely reported that the detection and comprehension of target speech in the presence of a noise masker depends on its position with respect to the masker on the horizontal plane [1–4]. An increase of the angular displacement of a masker and target has been shown to increase the effect of spatial unmasking [1,4]. Increasing the distance of a masker from a target has also been shown to increase the effect of spatial unmasking [4].

In the horizontal plane, spatial unmasking is primarily due to the better ear and binaural effects. The better ear effect relies on increased signal-to-noise ratios (SNRs) at one ear due to directional differences, producing an improved SNR at one ear due to interaural level difference (ILD) [4,5]. Binaural unmasking occurs as a result of different interaural time differences (ITDs) between the target and masker, resolved as phase differences between the ears at different frequencies, used to provide the brain with a less noisy representation of the auditory environment [6].

Previous studies prove that spatial unmasking persists in the median plane without azimuthal displacement. Martin et al. [7] showed the greatest mean percentage of correct target speech identification with the target at −50° elevation and masker at +50° elevation. This was a 27.5% increase from their worst condition, with both target and masker at −50°. McAnally et al. [8] showed that ITDs were not responsible for the significant spatial unmasking found in the median plane by removing the ITDs from head-related transfer functions (HRTFs) used in the experiment, finding the effect of spatial unmasking still to be present. These studies used speech-on-speech masking, and the resultant unmasking was largely informational due to the nature of the stimuli. Worley and Darwin [9] showed that subjects were able to track a speech's fundamental pitch with respect to a speech masker to aid differentiation from the masking speech source and provide further unmasking effects, which could contribute to the release of masking in speech-on-speech conditions if fundamentals were not matched. The current study sought to discover a speech reception threshold (SRT) at each masker elevation to compare the amount of spatial unmasking available at each location. Use of a speech-shaped noise masker helped to remove the informational aspects involved in speech-on-speech masking.

While previous research [7,8] suggests that the binaural influence on median plane spatial unmasking exists, albeit minimal, the range of tested elevation angles of the maskers was limited. Furthermore, the localization of sound in the median plane relies on spectral cues due to pinnae and torso diffractions (i.e., HRTF) [10–12], and therefore the spectral aspects of HRTF may also be related to spatial unmasking for speech intelligibility in the median plane. From this, it is hypothesised that if spatial unmasking for speech against noise operates in the median plane, it would be associated with the spectral energy difference between the target speech and masker noise at the ears, which would vary depending on the frequency notches produced at each elevation angle.

From this background, a subjective listening experiment was conducted to investigate the dependency of SRT in the median plane on the elevation angle of the masker, which varied with 30° intervals (−30°, 0°, 30°, 60° and 90°, with the target speech at 0°). The next section details the experimental method used. Section 3 presents the results of the listening test, which are discussed in Section 4 with objective analyses.

2. Method

Measuring SRT is a common method to determine the accuracy of identification of a speech source in the presence of a masker. SRT is defined as the lowest SNR at which 50% of the speech is correctly recognized [13]. The dB difference in SRT at different positions indicates the level of spatial unmasking. In the context of the present experiment, SRT is the difference of the level of a target speech signal, presented at 0° elevation in the median plane, to the constant level of a noise presented at −30°, 0°, 30°, 60° or 90° in the median plane to provide 50% speech identification.

2.1. Stimuli

The target speech signals were selected from a set of 720 high-context sentences from the Harvard IEEE corpus [14], which was also used by Shinn-Cunningham et al. [4] in their speech intelligibility experiments. Any sentences with pronunciations non-standard to British English were removed. The broadband speech recordings, all at 44.1 kHz sample rate and ranging between 1.3 and 2.6 s, were scaled for equal root mean square (RMS) value. An example sentence is as follows: “TAKE the WINDING PATH to REACH the LAKE”. The key words are displayed in capitals. There were five key words in each sentence.

Noise masker signals were generated uniquely for each target, as in Shinn-Cunningham et al. [4], a speech-shaped noise was used as the masker. An FFT was performed for each target and the phase components were randomised, producing noise with the same spectral content and length of the target. Shinn-Cunningham et al. [4] used a global noise masker with an average frequency content generated from the entire target database. In the current study, however, a speech-shaped noise masker was produced specifically for each target. This ensured consistency in target sentence evaluation whereby no single target sentence would be easier or more difficult to decipher with respect to its noise masker. This was considered to be important since the focus of the study was on the relative differences of SRT depending on the elevation angle of the masker, rather than an absolute SRT for each target.

2.2. Physical Setup

The listening test was performed in a double-walled ITU-R BS. 1116-compliant listening room (6.2 m × 5.2 m × 3.5 m) at the Applied Psychoacoustics Laboratory of the University of Huddersfield. The listening room has a short reverberation time (RT = 0.25 s) and a low noise level (NR = 12). Early reflections within 15 ms after direct sound generated by any loudspeaker used in this study were attenuated by minimum 22 dB for side walls and minimum 14 dB for the floor, which exceeded the requirement of the ITU-R BS. 1116 recommendation for critical listening room design. Therefore, it was considered that the room acoustics would have a minimal influence on the SRT measurement conducted in this study.

Genelec 8331A co-axial loudspeakers were mounted at −30°, 0°, 30°, 60° and 90° in the median plane. The distance between each loudspeaker and the listening position was 2 m, except for the 90° loudspeaker (1.4 m). All loudspeakers were level-matched and their room responses were equalised at the listening position using Genelec GLM 3.0 software.

The stimuli were reproduced via custom listening test software created in Max-MSP, through a Merging Technology (Puidoux, Switzerland) Horus audio interface. The target was always played from the 0° loudspeaker, while the masker was played from one of the five loudspeakers in each test trial. The target was presented initially at 44 dB L_{Aeq} and varied across the course of the test, while the level of the masker was held constant at 68.6 dB L_{Aeq}. This was calibrated using a DPA 4006A omni-directional microphone placed at a position corresponding to the subject's head.

2.3. Subjects

Six subjects participated in the experiment. They were postgraduate and final year undergraduate students of the Music Technology courses at the University of Huddersfield, aged between 21 and 27. All subjects reported to have normal hearing and were native English speakers. They all had previous experience with psychoacoustic listening tests.

2.4. Test Protocol

The SRT for each spatial configuration of target and masker was estimated using a two-down, one-up adaptive staircase procedure [15]. The five spatial configurations were randomly ordered for each subject. The adaptive procedure used in this experiment was largely based on the method used by Shinn-Cunningham et al. [4]. The test for each spatial configuration had two parts. In the first part, the subject was asked to confirm whether they had a correct identification of three keywords or more by clicking a "yes" or "no" box on the screen. If the response of the subject was "no", the level of the target was incremented by 4 dB and this process was repeated until the subject clicked "yes". Upon clicking "yes", the subject was asked to transcribe the presented sentence. Once this was completed, a correct transcript, with key words marked in red, was displayed on the screen below the transcription from the subject. The subject was asked to match the number of correct key words between transcripts and enter the number as a mark out of five (the maximum number of key words). The same speech source was used for this initial trial. Once "yes" was clicked, each further trial in the test used a unique speech sample randomly selected from the database, and the subject was asked to transcribe each sentence immediately following its presentation. The 4 dB increment was reduced to 1 dB following this point in the test. A correct trial was counted when 3 or more correct words (>50% identification) was recorded. For error marking, different suffixes were considered incorrect, but misspellings and homophones were considered correct. Two consecutive correct trials resulted in a 1 dB attenuation of the speech target (increasing the difficulty of the next trial). An incorrect trial resulted in a 1 dB increase of the speech target (decreasing the difficulty of the next trial). This continued until seven reversal points were completed. The average of the final five reversal points was used to estimate the SRT for each spatial configuration.

3. Result and Discussion

The results from the repeated measure ANOVA analysis, shown in Table 1, suggest a significant dependency of SNR on the masker elevation position ($F = 23.170$, $p = 0.000$), although this pattern is not linear, as can be observed in Figure 1. To examine which spatial configurations produced statistically different SRTs, paired-sample T tests with Bonferroni correction were conducted. The results are summarised in Table 2. The SNR for the −30° condition was significantly lower than any other condition ($p < 0.01$). It showed −3.5 dB of vertical spatial unmasking gain compared to the 0° condition ($p = 0.003$), whilst the 30° and 60° conditions had 1 dB ($p = 0.038$) and 1.6 dB ($p = 0.008$) gains, respectively. The 90° had a negative gain, but this was not statistically significant ($p = 0.446$).

Table 1. Results of repeated measure ANOVA (the sphericity is assumed based on Mauchly's test of sphericity ($p = 0.542$)).

Type III Sum of Squares	df	Mean Square	F	*p*
59.862	4	14.965	23.170	0.003

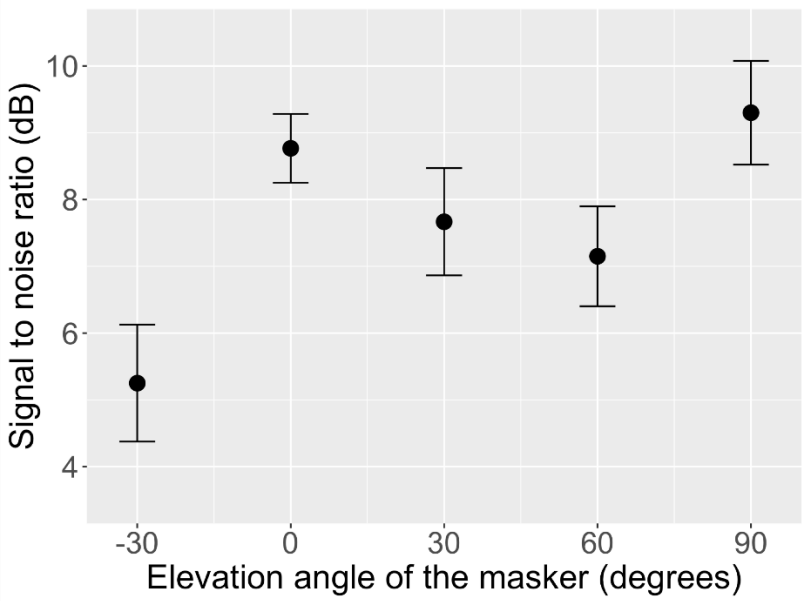

Figure 1. Signal-to-noise ratio (SNR) for 50% correct speech reception as a function of the elevation angle of the masker. The plots represent the mean average of SNRs obtained from six subjects and one standard error.

Table 2. Spatial unmasking gain on speech reception thresholds (SRTs): *p* values obtained from Bonferroni-corrected *t* tests. The gain is defined as mean difference of 0° from each masker position.

Masker Elevation (°)	Spatial Unmasking Gain (dB)	*p*
−30	3.5	0.003
30	1.1	0.038
60	1.6	0.008
90	−0.5	0.446

This result confirms the existence of spatial unmasking in the median plane. In order to obtain insights into the potential influence of spectral cue on the perceived results, HRTFs for the masker loudspeaker positions were analysed. As the original subjects were not available after the initial testing, head-related impulse responses (HRIRs) between −90° and +90° (at 30° intervals) in the median plane at 0° azimuth were taken from the SADIE II database [16] using the eighteen human subjects available. From this, an average HRTF was computed for each elevation angle by taking the mean value of the magnitude spectrum in decibels (number of FFT points = 8192, sampling frequency = 96 kHz). Since it has been shown that ITD has little influence on median plane SRT [8], and ILD is considerable in the median plane only above around 10 kHz [17], which is outside the important range of speech spectrum, only left ear signals were processed for this HRTF analysis. The current experiment included only one negative elevation (−30°) due to practical limitations in physical setup. Therefore, the subjective results presented here cannot generalise the consistency of the spectral notch influence on SRT at lower elevation angles. However, in order to provide further objective insight on the above hypothesis, −60° and −90° were also included in the spectral analysis.

Figure 2 plots HRTF differences of each tested masker position from the reference position 0°. Any positive value in the plots represents a greater energy compared to the reference, and vice versa.

The subjective results showed that the −30° masker condition produced the largest SRT gain of 3.5 dB among all conditions. Based on the delta HRTF plots, this might be explained as follows: Whilst the 30°, 60° and 90° conditions commonly display a high peak at around 8 kHz, it can be observed that the −30° and −60° conditions display a considerable notch at around 6 kHz, showing up to about 6 dB and 10 dB less energy than 0°, respectively. This is above the ranges that are considered to be most important for speech intelligibility (i.e., 2 to 4 kHz), but is still important for the reception of sibilant sounds and might therefore have been helpful in resolving confusion between words such as sat and that, which contain consonants with similar phoneme sounds. These phoneme sounds, called fricatives, display a peak between the range of 4 to 10 kHz depending on the sound pronounced [18]. Therefore, unmasking in this region could be greatly beneficial for the differentiation between words, thus benefiting SRT. In addition, the −90° condition displays the largest amount of attenuation against 0° above 1 kHz. This is likely to be due to a significant magnitude of shadowing effect by the human body when the source is placed directly below.

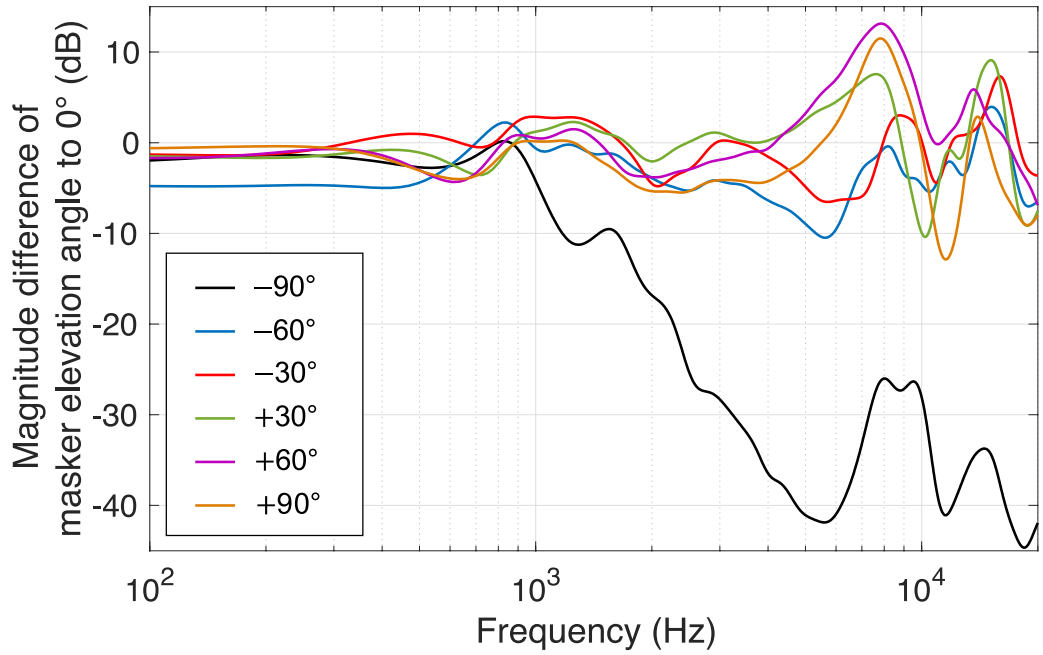

Figure 2. Head-related transfer function (HRTF) difference of each masker position to 0° elevation (reference) using average HRTFs of eighteen human subjects (SADIE II database [16]).

In addition, it is worth noting that similar studies on spatial unmasking in the median plane [3,4] reported SRTs around −6.4 dB and −9 dB respectively at 0° masker elevation angle. On the other hand, the current result demonstrates SRTs up to around +9 dB on average at 0°. This difference is likely due to the fact that the current study used a speech-shaped noise masker specific to each speech target sentence, rather than a noise masker averaged with respect to all sentences, thereby vastly increasing the difficulty of each trial. However, as mentioned earlier, this method ensures greater consistency for comparison between each vertical noise masker position, which was the main focus of the study.

4. Conclusions

This study demonstrated the existence of a spatial unmasking effect on the speech reception threshold (SRT) in the median plane. Among the five tested masker positions of −30°, 0°, 30°, 60° and 90°, the maximum gain of 3.5 dB was obtained at −30°, which was statistically significant. Gains at 30° and 60° were 1.1 dB and 1.6 dB, respectively, which were also significant. The 90° condition did not produce a significant gain. Based on the analyses of the HRTF magnitude differences of each masker position to 0°, it is suggested that a masker placed at a negative elevation in the median plane would

generally produce a larger spatial unmasking effect on SRT compared to one at a positive elevation, due to the energy reduction of ear-input signal at upper middle frequencies compared to 0°.

These findings are particularly relevant to three-dimensional (3D) audio mixing for speech clarity, particularly in film, or in any situation where background noise or other ambiences are presented in an environment where the impact on speech reception must be minimal. Providing a negative elevation to non-essential sounds could contribute to improved speech reception. In combination with the influence of visual stimuli on the perceived location of a sound source, this could be particularly effective in film to enhance dialogue intelligibility by negatively elevating other sources to floor-level loudspeakers in a 3D reproduction system.

Further work is required to confirm the consistent increase of spatial unmasking gain with increasing negative elevation in the median plane and to investigate spatial unmasking with three-dimensional displacements of a noise source. Furthermore, vertical spatial unmasking at azimuth angles other than 0° requires investigation.

Author Contributions: N.B. conducted the experiment, analysed the data and wrote the majority of the paper. H.L. supervised the project, contributed to the experimental design and data analysis and co-wrote the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the University of Huddersfield, Huddersfield, UK and Innovate UK, Swindon, UK (Grant Ref. 150175).

Acknowledgments: The authors are grateful to Siamäk Naghian and Aki Mäkivirta at Genelec (Iisalmi, Finland) for providing the loudspeakers used for the experiment. They also thank everyone who participated as a subject in the listening test.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Plomp, R.; Mimpen, A.M. Effect of the orientation of the speaker's head and the azimuth of a noise source on the speech-reception threshold for sentences. *Acoustic* **1981**, *48*, 325–328.
2. Durlach, N.I.; Colburn, H.S. Binaural phenomena. In *Handbook of Perception*; Academic: New York, NY, USA, 1978; Volume 4, pp. 365–466.
3. Bronkhorst, A.W.; Plomp, R. The effect of head-induced interaural time and level differences on speech intelligibility in noise. *J. Acoust. Soc. Am.* **1988**, *83*, 1508–1516. [CrossRef]
4. Shinn-Cunningham, B.G.; Schickler, J.; Kopčo, N.; Litovsky, R. Spatial unmasking of nearby speech sources in a simulated anechoic environment. *J. Acoust. Soc. Am.* **2001**, *110*, 1118–1129. [CrossRef]
5. Freyman, R.L.; Helfer, K.S.; McCall, D.D.; Clifton, R.K. The role of perceived spatial separation in the unmasking of speech. *J. Acoust. Soc. Am.* **1999**, *106*, 3578–3588. [CrossRef]
6. Glyde, H.; Buchholz, J.M.; Dillon, H.; Cameron, S.; Hickson, L. The importance of interaural time differences and level differences in spatial release from masking. *J. Acoust. Soc. Am.* **2013**, *134*, 147–152. [CrossRef] [PubMed]
7. Martin, R.L.; McAnally, K.I.; Bolia, R.S.; Eberle, G.; Brungart, D.S. Spatial release from speech-on-speech masking in the median sagittal plane. *J. Acoust. Soc. Am.* **2012**, *131*, 378–385. [CrossRef]
8. McAnally, K.; Bolia, R.; Martin, R.; Eberle, G.; Brungart, D.S. Segregation of multiple talkers in the vertical plane: Implications for the design of a multiple talker display. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Baltimore, MD, USA, 29 September–4 October 2002; pp. 588–591.
9. Worley, J.; Darwin, J. Auditory attention based on differences in median vertical plane position. In Proceedings of the 2002 International Conference on Auditory Display, Kyoto, Japan, 2–5 July 2002.
10. Roffler, S.K.; Butler, R.A. Factors that influence the localization of sound in the vertical plane. *J. Acoust. Soc. Am.* **1968**, *43*, 1255–1259. [CrossRef] [PubMed]
11. Blauert, J. Sound localization in the median plane. *Acoustic* **1969**, *22*, 205–213.
12. Algazi, V.R.; Avendano, C.; Duda, R.O. Elevation localization and head-related transfer function analysis at low frequencies. *J. Acoust. Soc. Am.* **2001**, *109*, 1110–1122. [CrossRef] [PubMed]

13. Determining Threshold Level for Speech. In *ASHA Practice Policy*; American Speech-Language-Hearing Association: Rockville, MD, USA, 1988; Available online: https://www.asha.org/policy/GL1988-00008.htm (accessed on 10 July 2020).
14. Rothauser, E.H.; Chapman, W.D.; Guttman, N.; Nordby, K.S.; Silbiger, H.R.; Urbanek, G.E.; Weinstock, M. Recommended practice for speech quality measurements. *IEEE Trans. Audio Electroacoust.* **1969**, *17*, 227–246.
15. Levitt, H. Transformed up-down methods in psychoacoustics. *J. Acoust. Soc. Am.* **1971**, *49*, 467–477. [CrossRef]
16. Armstrong, C.; Thresh, L.; Murphy, D.; Kearney, G. A perceptual evaluation of individual and non-individual HRTFs: A case study of the SADIE II database. *Appl. Sci.* **2018**, *8*, 2029. [CrossRef]
17. Carlile, S.; Pralong, D. The location-dependent nature of perceptually salient features of the human head-related transfer functions. *J. Acoust. Soc. Am.* **1993**, *95*, 3445–3459. [CrossRef] [PubMed]
18. Johnson, K. *Acoustic and Auditory Phonetics*, 3rd ed.; Wiley: Chicester, UK, 2011.