The role of head-related time and level cues in the unmasking of speech in noise

Barrie Edmonds and John Culling

School of Psychology, Cardiff University, South Glamorgan, UK, CF10 3YG

Summary

Two experiments investigated the effect of conflicting interaural time and level differences on the advantage of spatial separation between target speech and interfering sound. Target sentences were prepared using manipulated HRTFs from the Audis catalogue for azimuths of 0° and 60°. Speech reception thresholds (SRTs) were measured against interference with a 0° simulated azimuth, consisting of either Brown noise (-6 dB/oct. spectral roll-off) or speech. The results were consistent with independent contributions to intelligibility from time and level cues; their effects combine additively even when they indicate opposite directions for the target source. Experiment 1 compared the advantage of spatial separation cued by interaural time- and/or level-differences; each cue reduced SRTs by 3-4 dB, but their combined effect was 6-7 dB. Experiment 2 compared the advantage of spatial separation cued by combined cues that indicated the same or opposite-hemifield locations for the target speech. The combined effects of the two cues were indistinguishable, even when the conflicted with each other regarding the direction of the target speech. These results are inconsistent with a theory in which speech sounds are grouped by common location.

Introduction

In environments with multiple sound sources in different directions, one experiences a strong introspective impression that separate localisation of each source and attention to the direction of a particular target source is an essential aspect of listening. The assumption that it is necessary for listeners to attend to the location of a sound in order to perceptually separate it from interference was formalised in Bregman's (1990) account of auditory perception in which he describes the principle of spatial correspondence. This principle suggests that the spatial correspondence of sound elements promotes grouping into distinct auditory objects (i.e. sound localisation is a key step in the segregation process). Here we report data which suggest that this is not the case. Instead, they imply that sound-source localisation is irrelevant to the exploitation of binaural cues for speech intelligibility.

The principle of spatial correspondence is supported by the fact that perceptual segregation of a target sound from a concurrent interfering sound is better when the two sounds are spatially separated than when they are colocated. This effect is known as spatial unmasking, and, for speech, is often referred to as the binaural intelligibility level difference (BILD). The BILD is a measure of the gain in intelligibility due to binaural manipulations (such spatial separation of sound sources) in terms of the masked threshold of the target speech (in dB). Hirsh (1950) demonstrated that the BILD of target speech presented to listeners via loudspeakers varied with the relative azimuthal position of the loudspeaker presenting the target compared to the loudspeaker presenting the interferer. For instance, the masked threshold of speech was 4 dB lower when the target was presented from one location (e.g. left-hand-side loudspeaker) and the interferer from a different location (e.g. frontal loudspeaker) than when they were both presented from the same loudspeaker.

The physical location of a sound is not a cue directly available to listeners, but a perception based, for the most part, upon interaural time differences (ITD) and interaural level differences (ILD) between the waveforms arriving at the ears (Rayleigh, 1876, 1907). Correspondingly, differences in interaural timing and level of target and interferer contribute both to the BILD (Bronkhorst and Plomp, 1988), and to the localisation of each sound (Wigthman and Kistler, 19??). Of the two cues, ITD has been investigated the most extensively, as it is the dominant cue to sound localisation (Wightman & Kistler, 1992). The results of manipulating the ITDs of target and interferer have proven to be broadly consistent with the principle of spatial correspondence. That is, the BILD increases as the difference ITD between the target and interference is increased (Carhart, Tillman, & Greetis, 1969; Carhart, Tillman, & Johnson, 1968; Kock, 1950; Levitt & Rabiner, 1967; Schubert, 1956).

However, the relationship between the sound localisation and spatial unmasking is not as tight as one might hope. The dominance of ITDs over ILDs in sound localisation is not reflected by a dominant role for ITDs in the BILD. Bronkhorst and Plomp (1988) investigated the effects of spatial separation on speech intelligibility using head-related ITDs and ILDs and found that this was not the case. When they introduced a difference in ITD or ILD between a target sentence and noise interference the gain in intelligibility was roughly equal. Thus, although sound localisation is dominated by ITD, spatial unmasking is not. Other studies have shown that the BILD is greater in conditions that provide only diffuse localisation of the target voice (produced by a π-radian interaural phase shift), than in conditions that provide a clear image, generated by an ITD (Carhart et al., 1969; Schubert, 1956). Furthermore, Culling & Summerfield (1995) and Hukin & Darwin (1995) found that listeners appeared unable to group concurrent sounds according to common ITD as the principle of spatial correspondence would require. These studies suggest whilst spatial unmasking and sound localisation are both facilitated by ITD and ILD the former does not depend on the latter.

This paper investigates the role of sound localisation in the spatial unmasking of speech by dividing the target speech across different locations. In order to do this, our first experiment used an approach similar to that of Bronkhorst and Plomp to verify that ITD and ILD play equal roles in the separation of concurrent spatially separated stimuli. In the second experiment, ITD and ILD of the target were presented in conflict with one another such that each cue indicated a target sound source in different hemifield to the other. If ITD and ILD can be demonstrated to facilitate the unmasking of speech independently of one another when they indicate different spatial locations then sound localisation can play little role in the spatial unmasking of speech.

Experiment One: The contributions of ITD and ILD to the spatial unmasking of speech

Experiment One was a partial replication of the Bronkhorst and Plomp (1988) study, as it investigated the effects of time-only and level-only cues on the BILD. However, whilst they studied a range of azimuthal separations we have limited our investigation to just two: i) no spatial separation, and ii) 60° of spatial separation. SRTs in noise were measured over headphones. The interferer was Brown-noise or a male-talker presented with an ITD and ILD that were consistent with a sound source positioned straight ahead of the listener. There were four binaural conditions: i) baseline (the target was presented to the listener with an ITD and ILD identical to those of the interferer), ii) time-only (the target ITD indicated a sound source to one side of the listener whilst the ILD indicated a sound source straight ahead), iii) level-only (the target ILD indicated a sound source to one side of the listener whilst the ITD indicated a sound source straight ahead), and iv) time+level (the target ITD and ILD both indicated a sound source presented to one side of the listener).

Method

Participants. Two groups of 16 Cardiff University undergraduate psychology students were recruited and awarded course credit in return for their participation. All participants reported normal hearing and spoke English as their first language. Each participant was a naive listener (i.e. they had little or no previous experience in tests of auditory perception) and was tested only once in a session lasting approximately 45 minutes.

Stimuli. Sentences from the speaker CW from the M.I.T recordings of the Harvard Sentence List (IEEE, 1969) were used as target items. Interfering stimuli were either sentences from the speaker DA (again from M.I.T recordings of the Harvard sentence list) or a broadband noise. For the broadband noise, a white-noise was created digitally using the summed output of 16 consecutive pseudo-random numbers to produce each sample (see Klatt, 1980). The noise was then low-pass filtered through a 512-point FIR filter with a 6-dB/octave roll-off. As the waveform of the resulting broadband noise has a random-walk-like structure, it is called Brown-noise (after Brownian motion). Due to this spectral roll-off Brown-noise produces greater energetic masking for low-frequencies than higher-frequencies and roughly approximates the low-frequency emphasis of speech.

The Audis catalogue of HRTFs (Blauert et al., 1998) was used to produce HRIR files for the two ears at azimuths of 0° and 60°. Four HRIR pairs (i.e. one pair is comprised of a left-ear HRIR and a right-ear HRIR) were created for Experiment One with phase (*p*) and amplitude (*a*) components that were either congruent or incongruent with each other: p_0 a_0 (both the phase and amplitude components were consistent with a 0° source azimuth), p_{60} a_0 (the phase component was consistent with a 60° azimuth whilst amplitude was consistent with a 0° azimuth), p_0 a_{60} (amplitude was consistent with a 60° azimuth whilst phase was consistent with a 0° azimuth), and p_{60} a_{60} (both the phase and amplitude components were consistent with a 60° source azimuth)

These HRIRs were convolved with the target speech and interferer materials in order to produce four conditions (see Figure 1): i) **baseline** (the target speech was convolved with p_0 a_0), ii) **time-only** (the target speech was convolved with p_{60} a_0), iii) **level-only** (the target speech was convolved with p_0 a_{60}), and iv) **time+level** (the target speech was convolved with p_{60} a_{60}). The interferer in all cases was convolved with p_0 a_0.

Procedure Participants had their speech reception thresholds (SRTs) measured for each of the four conditions; the first group completed the task with the Brown-noise interference and the second group with the male-talker interference. The SRT is the masked level in dB of the target speech for a criterion level of performance. In this case, it was measured for the report of keywords from the target sentence with an accuracy of 50%. The measurement was implemented using the 1-up/1-down adaptive threshold method described by Plomp and Mimpen (1979). The BILD can be calculated by comparing the difference in SRT between conditions where the target and interference are convolved with different HRIRs and those conditions where they are convolved with the same HRIR.

Each condition was presented as a block of 10 trials. Each trial consisted of one target sentence and a concurrent interferer; the participant was informed that they were required to transcribe the target sentence and to score themselves on the number of correctly identified keywords. They repeated this procedure of transcribing and scoring for each trial in one condition before starting the next block of 10 trials (i.e. the next condition). This procedure was repeated until an SRT had been measured for the listener in all the experimental conditions.

For the first trial in each block (practice and experimental), the target speech was presented at a very low level compared to that of the interfering sound. A message presented via computer terminal, viewed through the booth window, then prompted the listener either to enter a transcript or to replay the first trial. If the participant replayed the stimulus the level of the target speech was increased by 4-dB. The first trial could be replayed in this way until it was loud enough to be judged partially intelligible by the listener (i.e. they felt they could hear approximately half the sentence). The participant then offered a transcript of the sentence, using the computer terminal, and the measurement entered a second phase in which the listener was given just one attempt to transcribe a fresh target sentence on each of the remaining trials for the current block.

In the second phase, the level of the target speech was adjusted up or down by ±2 dB in each trial based on the accuracy of the participant's previous transcript. If the participant reported transcribing 2 or fewer keywords the next trial target level was increased by 2 dB in the next trial otherwise the level of the target was decreased by 2 dB. After ten trials had been presented, the presentation levels used for the last seven trials and what would have been the eleventh trial was averaged (i.e. the mean value was taken) and used as a measurement of the speech reception threshold (SRT). The level of the interferer stayed the same throughout the entire block.

In order to eliminate order-effects the conditions were rotated around the different speech materials for successive participants. That is, each participant heard all the target/interferer speech materials in exactly the same order, only the order of the conditions was changed. Prior to commencing the experimental blocks all participants were given two practice blocks of monaural stimuli in order to familiarize themselves with the procedure. [bingo]

Results

Brown-noise interference Figure 2 shows the pattern of SRTs obtained from 16 participants who heard the target speech against Brown-noise interference. SRTs are highest for the baseline condition and lowest for the time+level condition. The time+level condition produces a binaural advantage of approximately 7 dB compared to the baseline condition. The time-only and level-only conditions gave intermediate thresholds.

A one-way repeated measures ANOVA was conducted for the four conditions and a significant effect was found (F(3,15)= 62.66 p<0.001). A Tukey test revealed no significant difference between the means of the time-only and level-only conditions. However, significant differences were found for: baseline vs. time+level (q=18.65, p<0.001), baseline vs. level-only (q=13.56, p<0.001), baseline vs. time-only (q=12.39, p<0.001), time-only vs. time+level (q=6.27, p<0.001), level-only vs. time+level (q=5.10, p<0.05).

Male-talker interference Figure 3 shows the pattern of SRTs obtained from the group presented with target speech and male-talker interference. SRTs are highest for baseline condition and lowest for time+level condition. The time+level condition produces a binaural advantage of approximately 5 dB over that of the baseline condition. Again, as with the Brown-noise interference, the time-only and level-only conditions gave intermediate thresholds.

Significant differences were found between the four conditions using a one way repeated measures ANOVA (F(3,15)= 14.15, p<0.001). All pair-wise comparisons were conducted with the Tukey test and revealed no significant differences for the time-only and level-only conditions or the level-only and time+level conditions. However, significant differences were observed for: baseline vs. time+level (q=9.00, p<0.001), baseline vs. level-only (q=6.15, p>0.001), baseline vs. time-only (q=4.60, p<0.05), and time-only vs. time+level (q=4.40, p<0.05).

Discussion

For the unmasking of speech with 60° of azimuthal separation between target and interferer Bronkhorst and Plomp reported a binaural advantage of 6 dB for their level-only condition, and 5.1 dB for their time-only condition. The results of Experiment One with Brown-noise interference compared well with their data as the binaural advantage over the baseline condition was 4.5 dB for the level-only condition and 4.1 dB for the time-only condition. The SRTs obtained with male-talker interference are lower than those obtained with the Brown-noise interference, however, the BILDs are comparable. Thus, when only ITD or ILD was made independently available each produced roughly equal contributions to the BILD in competing speech or noise for a spatial separation 0f 60°.

The BILD of the time+level condition in Brown-noise interference was 6.2 dB. This is larger than either the time-only or level-only condition BILDs. We assume that this increase in intelligibility is due to the combined benefits of unmasking due to ITD and ILD. However, the BILD observed in the time+level condition is smaller than would be predicted by simply summing the BILDs of the time-only and level-only conditions. Again, this is consistent with Bronkhorst and Plomp's data, as they showed that the gain in BILD for their free-field condition was not the sum of the BILDs due to ITD and ILD. Bronkhorst and Plomp concluded that the effectiveness of binaural unmasking by ITD is reduced in the presence of ILD. Such an argument is difficult to defend in this current investigation, as the difference in the BILDs for time-only and level-only conditions is negligible at 60° of azimuthal separation.

Experiment Two: The effect of conflicting head-related interaural time and level differences

If the effectiveness of ITD in binaural unmasking is reduced by headshadow then this poses a problem for accounts of perceptual separation based on spatial separation, as ITD is dominant in sound localisation not ILD. However, if ITD and ILD do not interact with one another then it should be possible to use both of these cues simultaneously, but independently to achieve spatial unmasking. In this second experiment, we investigated whether listeners could take advantage of both ITD and ILD even when they indicated different directions. This is an approach similar to that reported in the time-intensity trading literature (e.g. Deatherage, 1966; Deatherage & Hirsh, 1959; Hafter & Jeffress, 1968; Harris, 1960). The literature on time-intensity suggests that, for pure tones and impulsive stimuli, when the ITD and ILD of a stimulus are presented in opposition, such that each cue reflects a different sound source, the location of the consequent auditory image is affected in one of three ways. Firstly, the position can move towards an intermediate azimuth between those indicated by the opposing ITD and ILD cues (i.e. the two cues trade completely). Secondly, the auditory image can be diffusely localised (i.e. the two cues trade, but not completely). Lastly, the auditory image might split into a time-image and an intensity-image (i.e. the two cues do not trade at all; they enter localisation as distinct cues).

Thus, when target speech is presented with ITDs and ILDs in conflict with each other then we would expect that if spatial unmasking is dependent on sound localisation then this should be reflected in the BILD. In particular, SRTs should be higher in such a conflict condition compared those obtained in a condition with no conflict between ITD and ILD. It is suggested that if listeners can take advantage of ITD and ILD when they indicate different directions then segregation by sound localisation will be challenged.

Experiment Two tested the intelligibility of speech in noise using three binaural conditions: baseline, no-conflict and conflict. In the baseline condition the target was presented to the listener without spatial separation from the interferer. The listener was presented with a target having both ITD and ILD indicating a sound source 60° to the left of the interferer in the no-conflict condition. In the conflict condition the target was presented with an ITD indicating a sound source 60° to the left of the interferer, and an ILD indicating a sound source 60° to the right of the interferer.

If the SRTs for the conflict and no-conflict conditions are significantly different then the a directional hearing explanation of spatial unmasking will not be challenged, as one would predict conflict condition thresholds to be lower if grouping by sound location (i.e. comparable to a time-only or level-only BILD). However, if listeners are able to take full advantage of both ITD and ILD regardless of whether they point to the same direction or not then the conflict and no-conflict condition SRTs should be indistinguishable.

Method

Participants Two groups of 6 listeners were recruited. Again, as with Experiment One these listeners were Cardiff University undergraduate psychology students and were awarded course credit in return for their participation. All participants reported normal hearing and spoke English as their first language. Each participant had little or no previous experience in tests of auditory perception (and had not participated in Experiment One). All conditions were completed by each participant in a session lasting approximately ½ hour.

Stimuli The HRIRs created for Experiment One were reused here to generate the three conditions for Experiment Two (see Figure 4 for a schematic illustration). The interferer in all cases was prepared with the HRIRs for 0° azimuth (p_0 a_0). The baseline condition consisted of target speech convolved with the p0 a0 HRIRs (i.e. having phase and amplitude relationships that were consistent with a sound source at 0° azimuth). In the no-conflict condition the target was convolved with the p_{60} a_{60} HRIR (which specified phase and amplitude components consistent with a sound source at 60° azimuth). The conflict condition used the p_{60} a_{60} HRIR but had the phase and amplitude components in opposition to one another (the amplitude component of the left and right ear HRIRs were swapped around). This adapted HRIR, p_{60} a_{-60}, was convolved with the target speech to produce the stimuli for the conflict condition.

Procedure The SRT measurement for this experiment is slightly different to that described in for Experiment One. Listeners transcribed two blocks of 10 trials for each condition rather than just one block of 10 trials. Thus, six blocks of conditions were implemented (3 conditions × 2 blocks per condition) and they were counterbalanced and randomised for each listener as normal. The mean SRT was measured for each block as normal; the mean of the mean SRT for the two blocks for each condition was then calculated and is reported here. Otherwise, the SRT measurement did not differ in any other way from that described previously.

Results

Brown-noise interference Figure 5 shows the pattern of SRTs obtained for three conditions in Experiment Two for target speech presented concurrently with Brown-noise interference. SRTs were highest for the baseline condition, whilst the SRTs in the conflict and no-conflict conditions indicate improvements in intelligibility of about 6 dB.

A one-way repeated measures ANOVA was conducted for the three conditions and a significant effect was found (F(2,5)= 254.94, p<0.001). A Tukey test revealed No difference for the comparison of the conflict and no-conflict conditions. However, differences were found between the baseline and conflict conditions (q=28.83, p<0.001) and the baseline and no-conflict conditions (q= 26.31, p<0.001).

Male-talker interference Figure 6 shows the pattern of SRTs obtained for the three conditions in Experiment Two for target speech presented concurrently with male-talker interference. Again, SRTs were higher for the baseline condition than those measured for the conflict and no-conflict conditions which intelligibility improved by approximately 6 dB.

A one-way repeated measures ANOVA was conducted for the three conditions and a significant effect was found (F (2,5)= 11.95 p<0.005). A Tukey test revealed no difference for the comparison of the conflict and no-conflict conditions. However, differences were found between the baseline and conflict conditions (q=6.49, p<0.001) and the baseline and no-conflict conditions (q= 5.30, p<0.001).

Discussion

In Experiment Two, the listener was provided with target speech having both ITDs and ILDs different from those of the interferer, but with each cue indicating i) different directions (i.e. the conflict condition), or ii) the same direction (i.e. the no-conflict condition). It was predicted that if perceptual separation is based on grouping sounds by spatial location that the conflict condition BILD would be smaller than the no-conflict condition BILD. However, if the BILDs for these two conditions were indistinguishable then perceptual separation must have been possible without recourse to sound location, as listening to the location indicated by either ITD or ILD in the conflict condition would produce SRTs similar to those observed for the time-only and level-only conditions of Experiment One (i.e. higher than those seen in the no-conflict condition).

The BILD for the conflict condition was indistinguishable from that of the no-conflict condition. This suggests that, when separating target speech from a concurrent interferer listeners were able to take advantage of both ITD and ILD when they differ from those of the interference. Therefore, as listeners were able to take advantage of both cues even when they indicated directions in different hemi-fields it is proposed that the spatial unmasking does not seem to rely on the grouping of spatially corresponding sounds. Rather, binaural unmasking by ITD and ILD is responsible for the perceptual separation of speech in noise and not the consequent experience of sound localisation.

References

Bregman, A. (1990). *Auditory scene analysis: the perceptual organization of sound*. Cambridge MA: MIT-Press.

Bronkhorst, A. W., & Plomp, R. (1988). The effect of head-induced interaural time and level differences on speech intelligibility in noise. *Journal of the Acoustical Society of America, 83*(4), 1508-1516.

Carhart, R., Tillman, T. W., & Greetis, E. S. (1969). Release from multiple maskers: Effects of interaural time disparities. *Journal of the Acoustical Society of America., 45*(2), 411-418.

Carhart, R., Tillman, T. W., & Johnson, K. R. (1968). Effects of interaural time delays on masking by two competing signals. *Journal of the Acoustical Society of America., 43*(6), 1223-1230.

Cherry, E. C., & Bowles, J. A. (1960). Contribution to a study of the "cocktail party problem." *Journal of the Acoustical Society of America., 32*, 884.

Culling, J. F., & Summerfield, Q. (1995). Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay. *Journal of the Acoustical Society of America, 98*(2, Pt 1), 785-797.

Deatherage, B. H. (1966). Examination of binaural interaction. *Journal of the Acoustical Society of America., 39*(2), 232 249.

Deatherage, B. H., & Hirsh, I. J. (1959). Auditory localization of clicks. *Journal of the Acoustical Society of America., 31*, 486 492.

Hafter, E. R., & Jeffress, L. A. (1968). Two-image lateralization of tones and clicks. *Journal of the Acoustical Society of America., 44*(2), 563 569.

Harris, G. G. (1960). Binaural interaction of impulsive stimuli and pure tones. *Journal of the Acoustical Society of America., 32*(6), 685-692.

Hirsh, I. J. (1950). The relation between localization and intelligibility. *Journal of the Acoustical Society of America., 22*, 196-200.

Klatt, D. H. (1980). Software for a cascade/ parallel formant synthesiser. *J. Acoust. Soc. Am., 67*, 971-995.

Kock, W. E. (1950). Binaural localization and masking. *Journal of the Acoustical Society of America., 22*, 801 804.

Levitt, H., & Rabiner, L. R. (1967). Binaural release from masking for speech and gain in intelligibility. *J. Acoust. Soc. Am., 42*, 601-608.

Rayleigh, L. (1876). On perception of the direction of a source of sound. *Nature, 14*, 32-33.

Rayleigh, L. (1907). On our perception of sound direction. *Phil. Mag., 8*, 214-232.

Schubert, E. D. (1956). Some preliminary experiments on binaural time delay and intelligibility. *Journal of the Acoustical Society of America., 28*, 895 901.

Wightman, F. L., & Kistler, D. J. (1992). The dominant role of low-frequency interaural time differences in sound localization. *Journal of the Acoustical Society of America, 91*(3), 1648-1661.

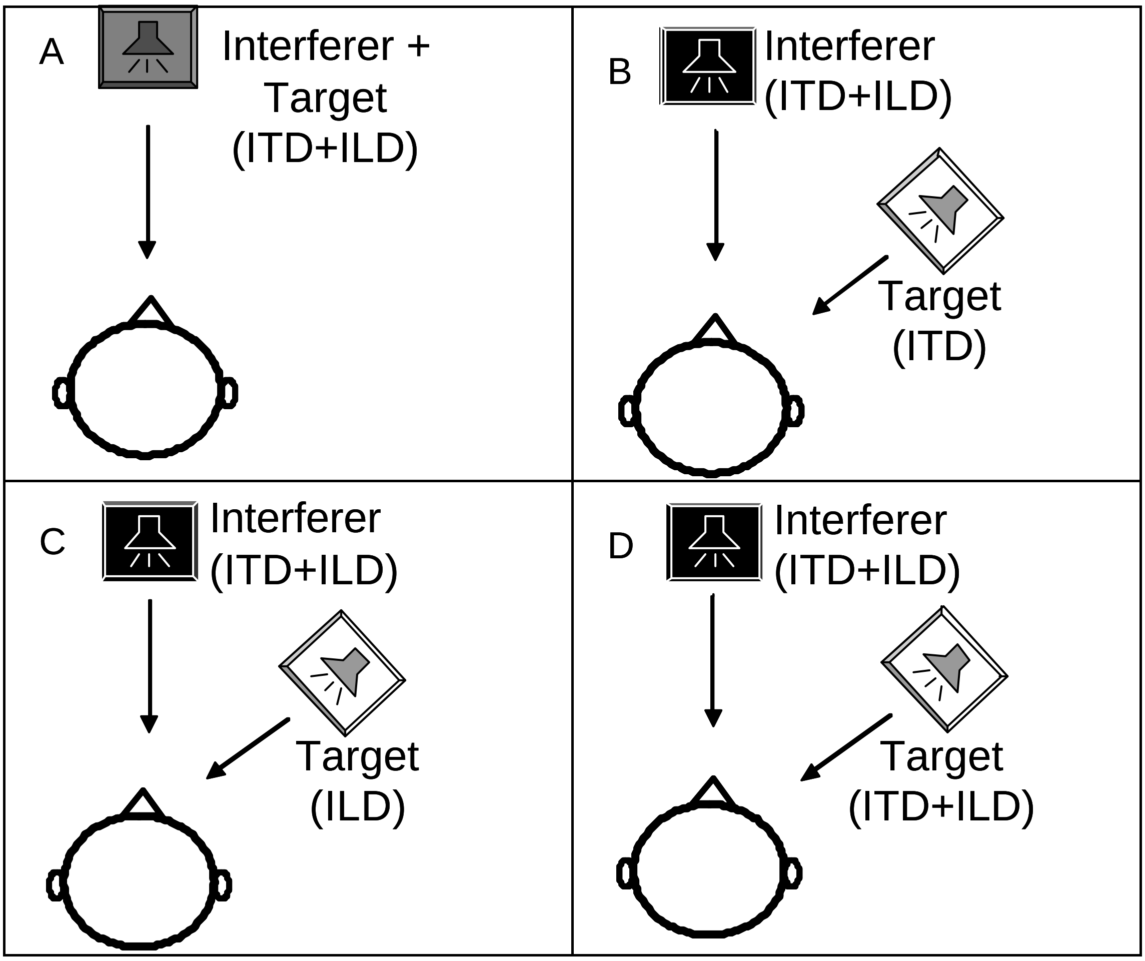

Figure 1: conditions for experiment 1

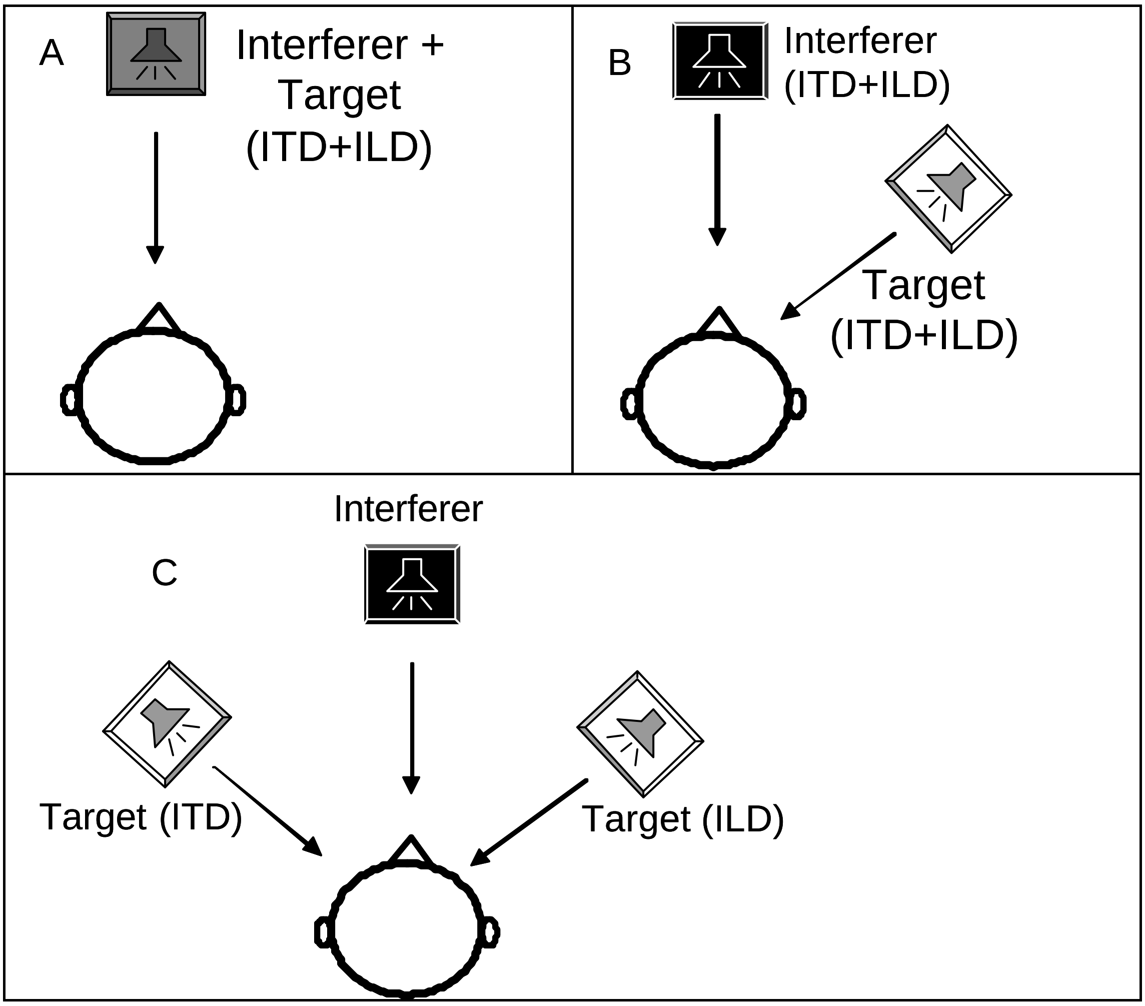

Figure 4: conditions for experiment 2