PERSISTENCE OF AUDITORY STREAMING PRESERVES RELEASE FROM MASKING

Kachina Allen[1], Simon Carlile[1], David Alais[1]and Karen Froud[2]
[1]Dept. of Physiology, University of Sydney, NSW. Australia 2106
[2]Dept. of Biobehavioural Sciences, Teacher's College, NY, USA, 10027
Correspondence: kachina_allen@yahoo.com

1. INTRODUCTION

It is widely recognised that spatially separating target speech from masking sound results in a listening advantage characterised by improved speech reception thresholds (SRTs) (eg., Freyman, Balakrishnan et al. 2001).

Yost (1991) argued that binaural and spectral cues assisted in the creation of separate auditory objects, allowing the listener to attend to one object and filter out the others. While these cues may be important, other factors are also relevant. Freyman et al (1999), and Driver (1996) showed that unmasking can be elicited using the illusion of spatial separation. This suggests that processes in addition to the salient location cues, may underlie a proportion of improvements in SRT due to spatial unmasking.

Where there are a number of concurrent sound sources, the association of spectral components with individual external sources probably involves firstly a short term grouping process and secondly a streaming of the grouped elements over time (Bregman 1994). In this experiment we aimed to test the contributions that each of these separate processes make to spatial unmasking.

2. METHOD

Subjects were seated, facing forward, in a sound attenuated, semi-anechoic chamber (size = 3.5 x 4.6 x 2.4m). Three Tannoy active loudspeakers were placed 1.3m away on the subject's audiovisual horizon.

Subjects had normal hearing, spoke English as a main language and included 4 females and 1 male (mean age 32 yrs). All subjects carried out 100 unrecorded practice trials of the separated, co-located and start separated condition and 150 trials of each condition.

Stimuli were generated and presented using Matlab and a Hammerfall multiface sound card at a sampling rate of 44100 and a volume of 57dB. Stimulus sentences were taken from the Coordinate Response Measure (CRM) corpus (Bolia, Nelson et al. 2000) and consist of an identifier in the first half (the call sign) and two target words (a colour and a number) in the second half. The target talker was identified by the call sign Baron and the subject's task was to identify the two target words in the presence of two masker talkers with different call signs and target words. Subjects entered the target words on a laptop. The signal to noise ratio of the target in relation to the maskers was varied randomly for each trial.

There were 4 conditions.

* **Co-located:** target and both maskers played from the central speaker.
* **Separated:** target played from central speaker, 1 masker played from each of symmetrically spaced speakers 30° from the central speaker.
* **Start Separated:** Target and masker start as in separated condition but collapsed to central speaker after 700 ms (just after the identifying call sign).
* **Start Co-located:** Target and masker start as in co-located condition but move to locations as in separated condition after 700 ms.

3. RESULTS

Subject SRTs were calculated using maximum likelihood generation of cumulative Gaussians at the 50% intelligibility level. Release from masking (RFM) was calculated as

RFM(condition) = SRT(condition) – SRT(co-located)

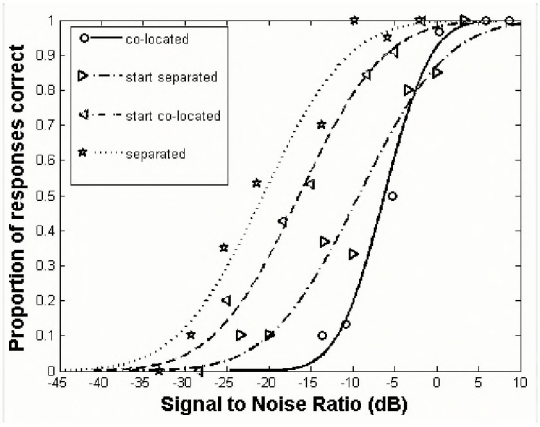

Figure 1 Representative psychometric functions (S4) for each condition. SRT calculated at 50% intelligibility

Bootstrapping was carried out with 500 repeats. Post-hoc t-tests ($\alpha = 0.05$) with Bonferroni corrections were carried out on bootstrapped data.

All subjects showed significant release from masking for all conditions (Table 1). Release from masking in the start co-located condition was much less than in the separated condition.

Subject	Start Separated RFM (dB)	Separated RFM (dB)	Start Co-located RFM (dB)
S1	4.2 ± 1.5	13.6 ± 1.2	15.3 ± 1.2
S2	2.1 ± 0.5	7.0 ± 0.8	5.2 ± 0.8
S3	3.6 ± 1.2	13.9 ± 1.2	13.7 ± 1.3
S4	3.5 ± 0.9	14.5 ± 0.9	11.2 ± 0.8
S5	4.5 ± 0.7	12.0 ± 1.1	9.4 ± 0.7
Mean	3.6	12.2	10.1

Table 1 Release from masking for conditions. RFM given ± standard deviations calculated from bootstrapped SRTs.

In the start co-located condition, subject SRTs were similar to those obtained for the separated condition. While there was a slight trend to reduced SRTs in the start co-located condition, this difference was significant in only one subject. This suggests that subjects may have been employing the strategy of simply listening to the central speaker location after the talkers separated rather than identifying and following the target talker.

To test this, a condition, "start co-located-all move", was added, in which subjects were forced to follow the target talker after separation. The target and masker co-located at the central speaker, then moved to locations where target and maskers were on different speakers after 700 ms, with the target randomly assigned to one of the three speakers.

Three subjects lost any release from masking in this condition. Where two subjects maintained significant release from masking, these subjects showed reduced performance (SRTs) on lateralised speakers indicating they were favouring the central speaker.

Subject	Start Co-located, all move RFM (dB)
S1	1.1 ± 1.3
S2	2.0 ± 0.8
S3	2.4 ± 1.2
S4	**2.3 ± 0.8**
S5	**5.9 ± 0.8**
Mean	2.1

Table 2 Release from masking for conditions. RFM given ± standard deviations calculated from bootstrapped SRTs. Those in bold show a significant ($\alpha < 0.05$) release from masking.

4. DISCUSSION

When target and masking voices are co-located, the listener relies entirely on cues such as voice characteristics (gender, tone, accent etc) to isolate the target talker. When separated, the listener can also use spatial information such as the binaural (inter-aural time and level differences ITDs and ILDs) and spectral cues to identify and maintain stream segregation of the talkers.

The spatial release from masking between the co-located and the 30° symmetrically separated condition (12.2 dB) is higher than that found in much of the literature (eg. 5dB, from Noble and Perrett 2002). The CRM corpus has the same carrier phrase for subject and maskers, with the same onset time. This synchronisation maximises both the energetic and informational masking. Other studies have used dissimilar discourse for targets and maskers. Spatial release from masking is often more marked with high information masking (eg. Noble and Perrett 2002; Hawley, Litovsky et al. 2004) and this may explain the higher level of unmasking in this study compared with previous studies.

Previous studies have disagreed over the relative contributions of voice cues and location cues in grouping and streaming (Mondor, Zatorre et al. 1998; Darwin and Hukin 1999; Edmonds & Culling 2005). In the current study, all subjects showed a significant level of unmasking in the start-separated condition. This may be due to the initial separation allowing the listener to more easily create and identify the target streams and thus detect voice characteristics of the target talker. While spatial cues disappear as the streams become co-located, these identifying characteristics can be used to continue to attend to the target or to filter out masker talkers.

The differences in unmasking between the separated (12.2 dB) and start-separated (3.6 dB) conditions may indicate the proportion of spatial unmasking which relates to streaming due to location cues and that which relates to separation allowing identification of the target. Further research will be required to discover whether varying the reliability of other cues will affect the reliance on spatial information.

REFERENCES

Bolia, R.S., Nelson, W.T. et al. (2000). A speech corpus for multitalker communications research. *JASA* 107(2): 1065-6.

Bregman, A.S. (1994) *Auditory Scene Analysis: The Perceptual Organization of sound*, MIT Press, Cambridge.

Darwin, C. J. & Hukin, R. W. (1999). Auditory objects of attention: the role of interaural time differences. *J Exp Psychol Hum Percept Perform.* 25(3): 617-29.

Driver, J. (1996). Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature* 381:66-68.

Edmonds, B. A. & Culling, J.F. (2005). The spatial unmasking of speech: evidence for within-channel processing of interaural time delay. *JASA* 117(5): 3069-3078.

Freyman, R, Balakrishnan, U. et al. (2001). Spatial release from informational masking in speech recognition *JASA* 109(5):2112-22

Freyman, R., Helfer, K. et al. (1999). The role of perceived spatial separation in the unmasking of speech. *JASA*, 106(6): 3578 - 3588.

Hawley, M., Litovsky, R. et al. (2004). The benefit of binaural hearing in a cocktail party: Effect of location & type of interferer. *JASA* 115(2): 833-843.

Mondor, T., Zatorre, R. et al. (1998). Constraints on the Selection of Auditory Information. *J Exp Psychol Hum Percept Perform.* 24(1): 66-79.

Noble, W. & Perrett, S. (2002). Hearing speech against spatially separate competing speech versus competing noise. *Percept Psychophys*, 64(8): 1325-36.

Yost, W. A. (1991). Auditory image perception & analysis: the basis for hearing. *Hear Res.* 56(1-2): 8-18.