Speech intelligibility in cocktail parties: the contribution of binaural cues to perceptual stream formation

Esther Schoenmaker and Steven van de Par
Carl von Ossietzky Universität Oldenburg, Institut für Physik - Akustik, 26111 Oldenburg, Deutschland
Email: esther.schoenmaker@uni-oldenburg.de

Introduction

Spatial separation of sources is associated with lower detection thresholds and improved speech intelligibility compared to non-separated sources at the same SNR [1]. Thus, binaural cues are assumed to enhance stream segregation in cocktail party listening, although the major cues are still likely to be monaural cues like pitch [2][3]. At the cognitive level, switching attention to a new location as well as listening to a target speaker at an uncertain location are associated with reduced speech intelligibility [4][5]. In this study we investigate whether continuity of interaural time delay (ITD) cues contributes to perceptual stream formation, in order to learn more about the importance of these cues in cocktail party listening. We used ITDs as the only binaural cue to avoid better ear effects.

Experiment

Three sequences of short utterances were presented simultaneously from different perceived directions in a headphone experiment. Listeners needed to attend to one target speaker and identify the interval that deviated from the rest of the sequence. Five different conditions were used, which differed in their continuity of ITD across utterances.

Material and method

Speech signals from four male and four female speakers were used. The utterances consisted of vowel-consonant-vowel (VCV) logatomes from the Oldenburg Logatome Corpus (OLLO) with seven different voiced middle phonemes. All fragments were rms-normalized and presented at a level of 65 dB SPL.

A graphical representation of a typical experimental trial is shown in Fig. 1. During each trial a set of VCV logatomes was employed that shared the same vowel. Three sequences of logatomes, each spoken by a different speaker, were presented simultaneously with synchronized onsets of logatomes. Prior to the sequence of logatomes, the keyword *OLLO* was used as a cue for the voice of the target speaker. Both interfering speakers were of opposite sex of the target speaker. The target sequence consisted of five instances of the same logatome and one that differed. The deviating logatome appeared in one of the last three intervals. The only restriction to the remaining logatomes was that same logatomes were never presented at the same time. Listeners were instructed to attend to the target speaker and report the logatome that was different by selecting it from a list. The score was calculated as the percentage correct answers.

Figure 1: Graphical representation of a typical experimental trial. The target speaker and target logatome are displayed in red. For more information, see text.

In table 1, the five ways in which ITDs were manipulated are shown. The ITDs were either kept constant or changed after each logatome in the sequence. Applying fixed or moving ITDs to the target speaker and/or the interferers resulted in four different measurement conditions (condition 2–5). Condition 1, in which all three speakers were presented with the same ITD, was added as a reference to estimate the effect of binaural masking release. Listeners were informed about the condition (i.e. fixed or changing direction) of both the target and maskers at the start of each new block of trials.

Table 1: Experimental conditions

Condition	Target ITD	Masker ITDs	Same ITD
1	Fixed	Fixed	No
2	Fixed	Fixed	Yes
3	Fixed	Changing	Yes
4	Changing	Fixed	Yes
5	Changing	Changing	Yes

The ITDs employed ranged from -700μs to $+700\mu$s. During measurement conditions (2–5), the three simultaneously presented speech signals were presented each with a different ITD. These ITD differences were restricted to multiples of 300μs.

Results

Eight normal-hearing subjects participated in the experiment. The mean overall performance was 62.8% correct. Scores per condition are graphicaly shown in Fig. 2. A one-way repeated measures ANOVA showed a significant main effect of condition ($F = 28.6$, $p < 0.001$).

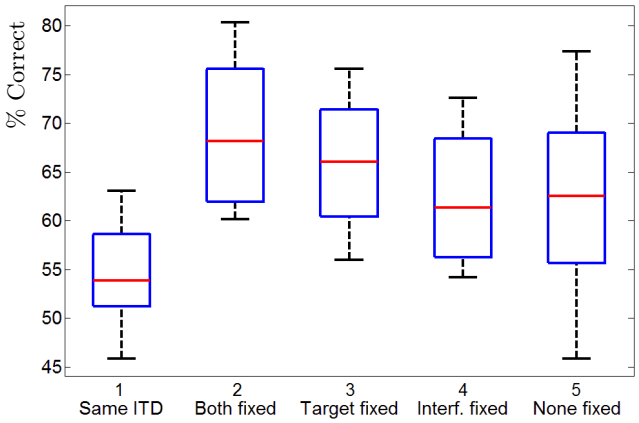

Figure 2: Boxplots of mean scores for all participants for each condition.

A set of planned comparisons revealed:

- A significant difference ($p < 0.001$) between conditions 2–5 where target and maskers were presented with different ITDs and the reference (condition 1) where all utterances were presented with the same ITD.
- A significant difference ($p = 0.002$) between the conditions with fixed target ITD and those with a changing target ITD.
- A significant difference ($p = 0.012$) between the two conditions with a fixed target ITD.
- No significant difference was found between the two conditions in which the target ITD was not fixed.

Post-hoc pair-wise comparisons with Bonferroni correction showed that the condition where none of the utterances was presented with a fixed ITD (condition 5) was the only condition for which performance did not differ significantly ($p = 0.110$) from condition 1 where all speech was presented with the same ITD.

Figure 3 shows the performance per condition when target intervals were sorted according to their ITD difference between target and nearest interferer. Only in the condition in which ITDs remained constant throughout the sequence, a significant linear increase ($p = 0.021$) of performance with ITD difference was found.

Conclusion

Results show an increased performance – indicating better streaming – for constant target ITDs during the course of a sequence. Constant interferer ITDs may offer a benefit as well, but this was seen only in those conditions where the target ITD was fixed. Besides the

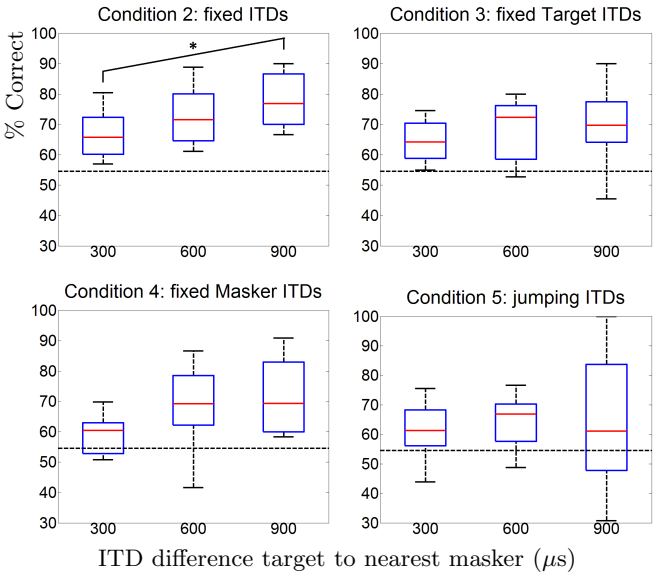

Figure 3: Boxplots of mean scores from all participants, for each minimal target-interferer distance in the target interval. The dashed line represents the overall mean score for condition 1, in which all speakers were presented with the same ITD.

positive effect on streaming, fixed ITDs seem to create a larger binaural advantage as indicated by a higher score for larger separations between target and interferers. This may be a matter of listening strategy. Since participants knew that (some) ITDs would be unpredictable they may have decided to rely on monaural cues only. On the other hand, spatial release from masking may have been counteracted by the uncertainty of where to direct spatial attention.

In conclusion, this study shows that perceptual streaming based on binaural cues is a component that contributes to speech intelligibility in complex acoustical settings.

References

[1] Bronkhorst, A.W.: The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. Acustica 86(1), 2000, 117–128

[2] Culling, J.F. and Summerfield, Q.: Perceptual Separation of Concurrent Speech Sounds - Absence of Across-frequency Grouping By Common Interaural Delay. J Acoust Soc Am 98(2), 1995, 785–797

[3] Darwin, C.J. and Hukin, R.W.: Auditory objects of attention: the role of interaural time differences. J Exp Psychol Hum Percept Perform 25(3), 1999, 617-629

[4] Best, V. *et al.*: Exploring the benefit of auditory spatial continuity. J Acoust Soc Am 127(6), 2010, EL258–EL264

[5] Brungart, D.S. and Simpson, B.D.: Cocktail party listening in a dynamic multitalker environment. Perception & Psychophysics 69(1), 2007, 79–91