Auditory Spatial Cuing for Speech Perception in a Dynamic Multi-talker Environment

Beáta Tomoriová and Norbert Kopčo
Katedra kybernetiky a umelej inteligencie FEI TU Košice, Slovakia
beata.tomoriova@gmail.com, kopco@tuke.sk

***Abstract*—An experiment was performed to measure how dynamic changes in a target talker location affect speech perception in a complex multi-talker environment. The listener's task was to report a number sequence spoken by the target talker, masked by four distractor talkers differing in their voices and spatial locations. The location of the target was either fixed during a trial (static condition) or randomly changing from one number to the next one (dynamic condition). The distractor talker locations were always randomized between numbers. A cue word sequence spoken by the target speaker with no distractors preceded each trial. The cue indicated the voice (dynamic condition) or the voice and location (static condition) of the target. A large decrease in performance was observed in the dynamic condition compared to the static condition. The position of the target number in a sequence, as well as the inter-word interval, only had a small effect on performance. The decrease in performance in the dynamic condition is either due to a cost of switching attention from one location to another, which can be large when attention is controlled by audition, or due to a change in listeners' strategy from selective focusing to one location (in static condition) to attempting to process all stimuli concurrently (in the dynamic condition).**

I. Introduction

Speech perception is a difficult process and attention plays an important role in it. In everyday-situations we face the problem of understanding what one person is saying while many other people are speaking at the same time. A typical example is the cocktail party problem [1] in which the listener has to select and follow one talker among a mixture of voices, noise, background music, and other distractors. Humans easily segregate the target voice from the mixture, attend to it, and suppress everything else. This ability seems to be quite simple and natural, but processes that underlie it are still not well understood. Many experiments were performed in this field that tried to find out which factors help segregate one "stream of speech" from another. One of the most important factors was spatial separation of speakers [2].

Several previous experiments studied spatial attention in a cocktail party problem and showed, that when the speaker location and voice are known, detection of a target message among masking messages improves [3, 4]. Recently, the interest has been expanded to performance in dynamically changing environments in which the listener does know ahead of time where to orient his/her attention. In such environments, cues used to orient attention can be provided by different modalities. Ozmeral et al [5] studied improvement in performance when attention is guided by a visual cue. The current study explores how listeners cope with such dynamically changing environment when only auditory cues are available.

II. Experiment

A. Motivation and hypotheses

The experiment presented in this paper is a pilot experiment for a study of spatial attention in cocktail party situations. The aim of this experiment was to find out how knowledge about spatial position of the speaker influences the ability to recognize his/her speech when only auditory cues are available.

Spoken digits were used as the speech stimuli and the listeners' task was to recall the sequence of digits spoken by a target voice in the correct order. Also, the effects of a digit position and inter-word-interval were evaluated.

Based on previous results, it was expected that:

1. the knowledge of the target speaker location will improve performance [4]
2. performance will be better at longer inter-word intervals, because the subjects have more time to "analyze" and process the input
3. understanding of speakers at side positions would be better, as seen in study [4]

B. Description of the experiment

Five subjects (3 women and 2 men aged 21-29) with normal hearing participated in this experiment. The subjects were native speakers of Slovak.

Speech stimuli were digits 1-9 spoken in English by five different male speakers and were taken from a TIDIGIT database [6]. Even though none of the listeners reported difficulty performing this task not in their native language, it is possible that this factor increased the difficulty of the task, and that, overall, better performance would be observed if a Slovak corpus was used.

The experiment was conducted in a regular quiet room in the Perception and Cognition Laboratory at TU Kosice. Speech stimuli came from five spatially separated loudspeakers, which were located approximately 120 centimeters from the subject at the level of the ears and were positioned at lateral angles of -40˚, -20˚, 0˚, 20˚, 40˚.

Under the middle loudspeaker, there was a computer screen displaying instructions to the subjects. The subjects were asked to fixate their sight at the mark "x" that appeared at the monitor during the presentation of stimuli.

At the beginning of each session, the subject was seated on a chair facing the loudspeakers, and given instructions about the experiment. His/her task was to listen to the

sequence of four random digits spoken by the voice specified by a cuing sentence that preceded the target (target voice). This target voice was during the presented sequence accompanied by distractor (masking) voices speaking digits (different voices) from remaining loudspeakers. All voices were male. Five voices were selected out of 15 to enhance identification based on voice.

The target-voice cue presented at the beginning of each trial consisted of 4 random numbers (similar to the targets). Then, four target spoken digits were presented at a constant rate, accompanied by masking distractor words. The distractors were presented synchronously with each target word from four other speakers. Also, the target and distractors were always different numbers and spoken by different talkers.

Another parameter examined in the experiment was the duration of the pause between the digits. Three different IWI (inter-word intervals) were used: 0, 500 and 2000 miliseconds. The IWI was kept constant within a trial, including both the cue and the target sequence. As not all recordings had the same duration, the IWI represented the duration between the end of the longest recording in one segment of sequence and start of the second segment.

The experiment consisted of two conditions labeled here as "environments", each represented in one block of the experiment and defining the type of the block:

1. Static environment – during one trial the target voice was presented from a fixed loudspeaker – indicated by the cue
2. Dynamic environment – target voice was presented from a sequence of loudspeakers randomly changing from one word to the next one, both in the cuing and in the target part of the sequence.

In the static blocks the cue provided information about the position of the target voice to which the subjects could attend. In the dynamic blocks the cue did not provide information about position, so the listeners first had to recognize the target voice and only then could direct their attention to the voice and to the loudspeaker from which it was spoken.

One block (static or dynamic) contained 15 trials, all combinations of a position of a target voice (5) and IWI (3). During one block each of the 5 available voices was chosen as a target with equal probability.

The whole experiment consisted of ten sessions. One session included one static block and one dynamic block in random order.

C. *Experimental procedure*

At the beginning of the experiment, the subjects did one training session to familiarize themselves with the experimental procedures and with the different voices used in the study. At the beginning of each experimental block, the subject was informed about the type of block (static or dynamic). Each trial started by providing information about the IWI used in the trial. Then the cue was presented using the same voice and the same IWI as in the target sentence. In a static block, the target voice came from the same loudspeaker during the introduction and the sequence. In dynamic blocks, the target came from a random loudspeaker at each segment of the cue and target sequence. The interval between the cue and the target sentence was the same as the IWI between the target words. When the sequence ended, the subject entered four digits representing his/her best guess of the sequence produced by the target voice.

D. *Scoring*

Results were scored by two different methods:

1. whole sequence – a response was counted as correct only if all four digits were reported correctly and in the right order (corresponding to chance performance of $(1/9)^4$),
2. digit-by-digit - each digit scored separately (corresponding to chance performance of 1/9).

Answers were binned by the type of environment (static/dynamic), position, the IWI, and the time segment (only when scored by the second method). Only digit-by-digit scores are reported here.

Percentage of correct answers for various combinations of these conditions were calculated for each subject and then averaged across subjects to get final results.

III. RESULTS

Figure 1 shows percent correct scores for each environment when using digit-by-digit scoring. The graph shows data for each subject and the across-subject mean. Performance in the static environment reached 52%, and was much better than in the dynamic environment (24%). However, even in the dynamic environment the performance was well above chance, indicating that the listeners could perform the task.

Figure 2 shows percent correct scores as a function of the time segment of a sequence. The effect of position of a digit within a sequence was very small.

Similarly, the dependence of performance on the length of the inter-word interval is very small (Figure 3).

Figure 4 shows performance as a function of the location of the target speaker. In the dynamic environment, performance is almost equal at all locations except for slight increase at the lateral locations. In the static environment, this dependence is enhanced. When the target voice was presented from loudspeakers positioned at –20˚, 0˚ or 20˚, performance varied between 40-47%, but when it spoke from the loudspeakers at the -40˚ or 40˚, as many as 64% of target digits were identified correctly.

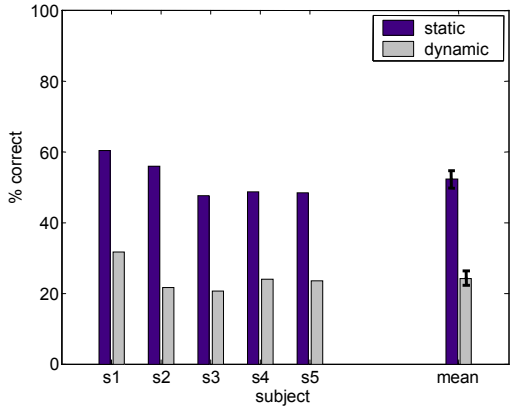

Figure 1. Percent correct scores. Left part of figure shows data for individual subjects and right-most bars show across-subject mean. Errorbars represent the standard error of the mean.

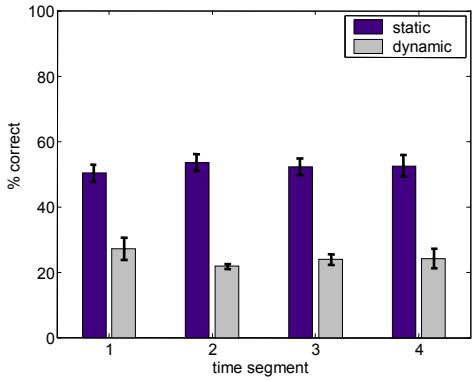

Figure 2. Percent correct scores as a function of a time segment (pooled over IWI, positions and subjects). Errorbars represent the standard error of the mean

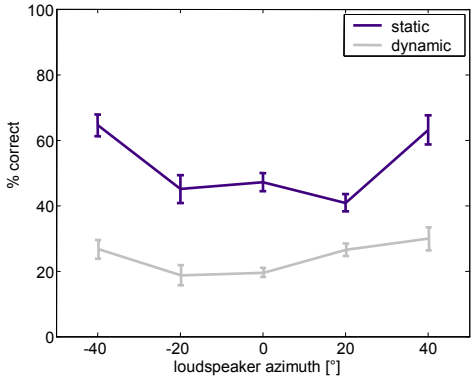

Figure 4. Percent correct scores as a function of location defined by loudspeaker azimuth (pooled over time segment, IWI and subjects). Errorbars represent the standard error of the mean.

IV. Discussion

The results of the current experiment indicate that *a priori* knowledge of the spatial location of a speaker improves the ability to process utterances spoken by a given voice way beyond improvement provided when voice is the only cue.

Increasing the delay between words does not influence performance this task (Fig. 3), even though in a similar study, in which cuing was visual, improvement was observed [5]. This suggests that a different strategy is used to orient attention when auditory cue is used.

Similarly, Ozmeral [5] found an improvement in performance for digits presented in later segments compared to the early segments, while no such improvement was observed here (Fig. 2). It is likely that the improvement in Ozmeral study resulted from gradual focusing of listener's attention, which in Ozmeral study started with the presentation of the first stimulus. In contrast, in the current study the listeners heard a four-word cue sequence presented from the target location using the target word. Thus, they could orient their attention before the first target word was presented.

Performance depended on specific position particularly in the static environment. When the target was presented from the left-most or the right-most loudspeaker, then it was recognized much better than at other locations. Similar results were observed in the study of Best et al. [4], which used a similar spatial setup. Best et al. suggested that this dependence comes from an improvement in the better-ear target-to-masker energy ratio (TMR) at the lateral positions. An alternative explanation is that the improvement comes from the fact that there is a neighboring distractor only on one side of the target for the extreme targets, while there are two neighboring distractors (one on each side) at the remaining locations. Thus, if the attentional filter is broader that the azimuthal separation of the loudspeakers, then interference from only one distractor enters the filter at the outermost locations, whereas two distractors interfere with the target at the central locations. It is more difficult to explain why similar benefit was not observed in the dynamic environment. It is possible that in this condition the strategy adopted by the listeners was to try to focus on the whole scene, instead of reorienting from one location to the next one. In such a case, the dependence on location would be expected if it was caused by a change in the TMR, but possibly not if it was due to the breadth of the auditory filter. Thus, this result can be interpreted as suggesting that the spatial dependence is caused by attentional filtering, not by TMR energy effects.

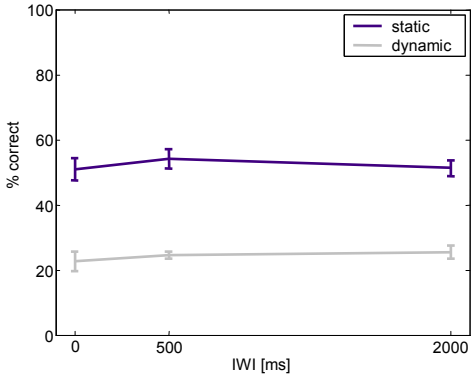

Figure 3. Percent correct scores as a function of IWI (pooled over time segments, positions and subjects). Errorbars represent the standard error of the mean.

Further experiments and analysis is necessary to provide a full interpretation of the current results, Specifically, it is not clear whether the poor performance in the dynamic condition comes from the listeners' inability to localize the target, to re-orient their attention to the target, or from a change in the strategy used by the listeners in the dynamic task.

Acknowledgment

Authors would like to thank students who helped to build experimental setup and who participated in the experiment as subjects.

This work was supported by grant from Slovak Scientific Grant Agency VEGA 1/3134/06.

References

[1] Yost, WA. (1997) "The cocktail party problem: forty years later." In: Binaural and spatial hearing in real and virtual environments (Gilkey RH, Anderson TR, eds), pp 329-347. Mahwah, NJ: Erlbaum.

[2] Cherry, E. C. (1953), "Some experiments on the recognition of speech, with one and with two ears." *Journal of Acoustical Society of America* 25, 975--979.

[3] Kitterick, P.T., Summerfield, A.Q. (2007) "The role of attention in the spatial perception of speech", Association for Research in Otolaryngology, Abstract #749

[4] Best, V., Ozmeral, E., Shinn-Cunningham, B., "Visually guided Attention Enhances Target Identification in a Complex Auditory Scene," in *Journal of the Association for Research in Otolaryngology,* vol. 8, 2, June 2007, pp. 294-304(11).

[5] Ozmeral, E., Best, V., Kopco, N., Mason, CH., Kidd, G., Shinn-Cunningham, B., (2008) "Dynamic aspects of auditory spatial attention", Association for Research in Otolaryngology, Abstract

[6] Leonard, R. G., (1984). A Database for Speaker-Independent Digit Recognition. Proc. ICASSP 84, Vol. 3, p. 42.11, 1984.