The relative contributions of better ear listening and binaural masking level differences in a cocktail party: Experiment and model predictions

Esther Schoenmaker[1], Thomas Brand[2] and Steven van de Par[1]
[1] Acoustics group, [2] Medical physics group, Carl von Ossietzky Universität Oldenburg, 26111 Oldenburg, Deutschland
Email: esther.schoenmaker@uni-oldenburg.de

Introduction

Psychoacoustic experiments with collocated and spatially separated speech sources have revealed a higher speech intelligibility for separated than for collocated speakers [1]. This phenomenon is known as spatial release from masking. The results are similar to the detectability of a tone in noise that has a different interaural phase than the tone. This has led to the idea that equalization-cancellation (EC) type models which have been proven successful in explaining masking release for tone in noise detection [2] can also be applied to predict speech intelligibility. An example of a successful implementation of this theory is the binaural speech intelligibility model (BSIM) by Beutelmann and Brand [3]. On the other hand, unlike stationary noise, speech signals are typically strongly modulated both in the time and frequency domain. This property provides listeners with the opportunity to collect glimpses of information about the target speech [4]. Interaural difference cues may then facilitate the segregation of glimpses originating from different directions. Since a substantial amount of glimpsed information will be available at relatively high signal-to-noise ratios (SNR), the question rises whether the contribution of simultaneous interaural difference cues, that in principle allow for a binaural release from masking, is still relevant for speech intelligibility. In this paper an answer to this question is sought with the help of speech stimuli that contain reduced interaural difference cues, in combination with a normal spatial image. Model predictions for these stimuli will be compared to psychoacoustic data.

Experiment

Material and method

The speech material consisted of vowel-consonant-vowel logatomes from the OLdenburg LOgatome Corpus (OLLO, [5]). Sequences of logatomes spoken by three different speakers were presented simultaneously with synchronized logatome onsets. A typical experimental trial is shown in Fig. 1. The voice of the target speaker was signaled to the listeners by the keyword "ollo" directly prior to the start of each trial. The target speech consisted of a sequence of six logatomes, of which five were identical and one (presented during one of the last three intervals) was different. The interfering speech consisted of random logatomes spoken by two speakers of different sex. The listeners' task was to pay attention to the target voice and to select the deviating logatome from a closed set of logatomes which differed only in their middle consonant. The performance was expressed as the percentage correct answers. Eighteen normal-hearing subjects participated in the experiment.

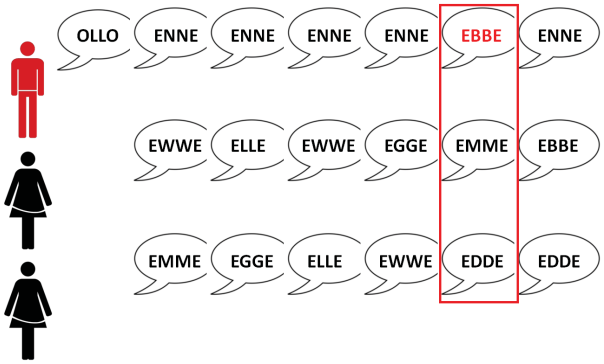

Figure 1: Graphical representation of a typical experimental trial. The target speaker and target logatome are displayed in red. The keyword "ollo" serves as a cue to the target speaker. The red box indicates those parts of the speech signals that served as an input to the binaural speech intelligibility model.

Conditions

Two different stimulus types were used that were spatialized using different methods. Stimuli of the first type were filtered with anechoic head-related transfer functions (HRTF) recorded with an artificial head and torso simulator. Accordingly, natural interaural cues corresponding to sources located in the frontal horizontal plane were imposed on the signals. The second type of stimuli were equal local azimuth (ELA) stimuli that have been described in detail in [6]. These stimuli were derived directly from HRTF-filtered signals. However, their interaural properties were manipulated within time-frequency units of 12 ms in length and one ERB wide. In each time-frequency unit the three speech sources effectively became collocated. The positions of collocation were chosen according to the strongest source within the particular unit. Due to the sparse and modulated character of the speech signals, any of the three speaker positions will dominate the mixture at some time-frequency units and any position information will thus be present in the final signal. Overall, the spatial quality of the original mixture was closely preserved, but any instantaneous interaural difference that could give rise to a binaural masking level difference were excluded from the stimuli.

Both signal processing methods (HRTF and ELA) were combined with collocated and spatially separated virtual

speech sources, leading to 2 x 2 = 4 different conditions. Note that the ELA manipulation will not affect the already collocated stimuli, except for the potential introduction of processing artifacts. All stimuli were presented over headphones at a level of approx. 65 dB SPL. Each participant listened to 168 trials in each of the four conditions.

Binaural Speech Intelligibility Model

A detailed description of the BSIM can be found in [3]. In short, the model calculates two monaural SNRs per frequency band based on the long-term spectra of the target and interfering speech signals. Furthermore the model calculates a binaural SNR exploiting the interaural differences of the stereo signal using EC processing. The speech intelligibility index (SII, [7]) is then calculated based on the optimum of these three model output SNRs in each frequency band. The model was used to calculate SII predictions for only the target intervals of the stimuli that had been presented to a random participant in the experiment. The underlying assumption is that the listeners were well able to detect the proper deviant interval and that speech intelligibility predominantly depended on the combination of the utterances and source directions within this target interval.

Results

The results of the experiment and the model predictions are shown together in Fig. 2. The experimental data show a spatial release from masking for both stimulus types. The ELA stimuli, however, led to a somewhat smaller masking release than the HRTF-filtered stimuli. The data points for the two HRTF conditions were used to map the SII data from the model to the mean percentage correct answers in the experiment. As can be seen from Fig. 2, the predictions for the collocated ELA stimuli fit well to the data. The SII prediction (and thus the corresponding masking release) for the separated ELA condition is smaller than its matched experimental result.

Discussion

Both the psychoacoustic data and the model predictions show a spatial release from masking for the ELA stimuli. These stimuli contain separate interaural difference cues for each speaker, a feature that is necessary to achieve spatial release from masking according to the EC-theory. The interaural difference cues, however, were never present simultaneously, but existed only between time-frequency units. From the masking release that is observed in the psychoacoustic data it appears that the differential place cues need not be present simultaneously, but can be combined over time by the auditory system to achieve a benefit of spatial separation. This is indicative for a glimpsing model of speech intelligibility rather than the processing of binaural masking level differences. The reduced release from masking for the ELA stimuli as predicted by the BSIM suggests that the model suffers more from the removal of simultaneous masking release cues than human listeners do. Nevertheless, the BSIM can still predict part of the observed masking release. This can be traced back to a long-term SNR improvement that results from a stronger suppression of the interfering speech relative to the target speech.

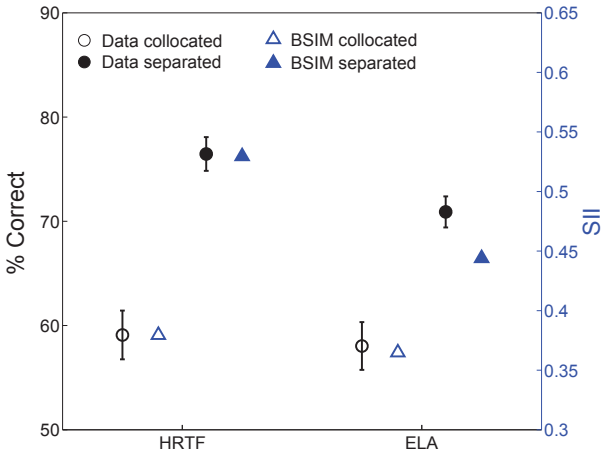

Figure 2: Data and model predictions for the original HRTF-filtered and ELA-manipulated stimuli. Circles show the mean scores and standard errors of the experiment (left axis). Triangles show SII predictions by the Binaural Speech Intelligibility Model (right axis). Open symbols indicate collocated sources, whereas closed symbols represent separated source conditions.

References

[1] Peissig, J. and Kollmeier, B.: Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners. J.Acoust.Soc.Am. **101** (1997), 1660-1670

[2] Durlach, N.I.: Equalization and cancellation theory of binaural masking-level differences. J.Acoust.Soc.Am. **35** (1963), 1206-1218

[3] Beutelmann, R., Brand, B., and Kollmeier, B: Revision, extension, and evaluation of a binaural speech intelligibility model. J.Acoust.Soc.Am. **127** (2010), 2479-2497

[4] Cooke, M.: A glimpsing model of speech perception in noise. J.Acoust.Soc.Am. **119** (2006), 1562-1573

[5] Wesker, T. et al: Oldenburg Logatome Speech Corpus (OLLO) for Speech Recognition Experiments with Humans and Machines. Proc. of Interspeech (2005), 1273-1276

[6] Schoenmaker, E. and Van de Par, S.: Auditory streaming in cocktail parties and the extent of binaural benefit. POMA **19** 050157 (2013)

[7] ANSI S3.5-1997: Methods for the calculation of the speech intelligibility index. American National Standards Institute, New York (1997)