EC-processing and Glimpsing in Cocktail Party Situations

Sarinah Sutojo, Esther Schoenmaker and Steven van de Par
Carl von Ossietzky University, Acoustics Group, 26111 Oldenburg, Germany
Email: sarinah.sutojo@uni-oldenburg.de

Introduction

In multiple-talker situations or in the presence of masking noise, speech intelligibility can be improved by spatially separating the source of the target signal from that of the masker. This advantage in intelligibility is generally referred to as Spatial Release from Masking (SRM) and can be observed in the difference of speech reception thresholds (SRTs) for collocated and separated sources. To gain a deeper understanding of the mechanisms that lead to this improvement in speech intelligibility, different auditory models have been suggested [3], [4]. The present study is concerned with the contributions of two such explanatory approaches to SRM in cocktail party situations.

EC-processing and Glimpsing

One of the two mechanisms, which were considered in the experiment, relies on the exploitation of simultaneous binaural cues. Commonly known are experiments for tone detection in noise, in which the tone becomes more detectable when target and masker signal have different sets of binaural cues. This improvement in tone detectability for binaural conditions is known as the Binaural Masking Level Difference (BMLD). It is obtained by comparing the masked threshold in the case of same phase and level relationships for target and masker at the two ears, with the case of interaural cue differences between the two signals [1]. A model typically associated with the explanation of BMLDs is the Equalization Cancellation (EC) Model [2], which has been further extended to be applicable to speech intelligibility [3]. According to this model, the raise in detectability or intelligibility is due to an attenuation of the noise in the internal representation of the total signal. This attenuation is achieved by equalizing phase and level of the noise at both ears and then canceling by subtraction of both ear channels. The remaining signal then contains mainly target signal and a residual error, which leads to an improved SNR. From behavioural data it is known that the binaural improvement in detectability of a target signal is only obtained when simultaneous interaural cue differences are available to the listener. Furthermore, it is assumed that this mechanism is effective only in spectro-temporal regions with negative local SNRs; for positive SNRs, monaural processing already allows the detection of the target signal.

A different approach to understand speech intelligibility in a complex auditory scenario is the exploitation of so called "glimpses". Due to speech being highly modulated in time and frequency, in certain spectro-temporal regions target energy rises above the masker energy. Such sparsely distributed regions are referred to as glimpses and are being collected by the listener. According to Cooke [4], these glimpses could also be defined as regions in which the local SNR rises above a certain value such as -2 dB or -5 dB, suggesting that slightly negative SNRs are still useful to the listeners. Following this second strategy the listener draws the relevant information for speech intelligibility from these fragments in which the target signal is minimally impaired by the masker.

The basic difference in the two considered approaches to explain SRM lies in the type of cues from which the listener draws the relevant information about the target signal. While the first approach is exploiting simultaneous interaural cue differences between target and masker signal and effectively operates in regions of moderately negative local SNRs (i.e. SNR<-5 dB), the glimpsing approach assumes a main contribution of regions with favorable local SNRs (i.e. SNR>-5 dB), without the necessity of simultaneous interaural differences.

An experiment was conducted to compare contributions of both mechanisms to speech intelligibility, applying a stimulus manipulation that eliminates the possibility to perform BMLD-like processing but allows glimpsing. SRM was measured in the presence of different maskers that vary in the amount of glimpsing opportunities they offer. Furthermore, a measure is introduced that accounts for the amount of available glimpses in a masked speech segment.

Experiment

The guiding question to the conducted experiment was, how large the benefit from simultaneously available binaural cue differences would be. Based on the outcome, it should be possible to deduce information about the contributions of a BMLD-like mechanism. In addition to that the question of how well the listeners' performance could be explained when assuming a glimpsing approach is investigated.

Stimulus Manipulation

The concept of the experiment was to distinguish contributions of both mechanisms by applying a stimulus type that thwarts one type of processing, allowing only the other mechanisms to operate. In this case the BMLD-like processing was prevented by removing simultaneous binaural cue differences, while maintaining the possibility of glimpsing. The stimulus manipulation used in the experiment is referred to as Inferior Speech Elimination (ISE).

The central idea of the manipulation is to eliminate simultaneous binaural cues for both sources by eliminating the inferior signal, i.e. the signal with lower local SNR, within each time-frequency unit. Hence, only one signal, the dominant signal, remains present during one instance of time and frequency. Informal listening showed that a natural spatial impression is maintained, making it difficult for the listeners to distinguish the manipulated stimuli from the unaltered ones. A diagram of the signal processing that was done in the ISE manipulation is given in Figure 1.

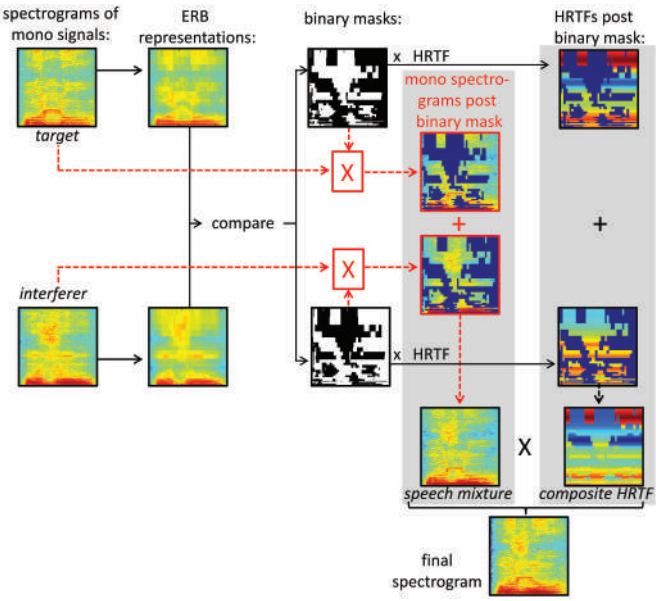

Figure 1: Diagram of ISE-stimulus manipulation.

To simplify the chart, the signal processing steps are shown for one ear channel only. On the very left, spectrograms of exemplary mono speech samples are shown for target and interferer. As a first step, these spectrograms are transferred into an Equivalent Rectangular Bandwidth (ERB) representation. This representation is a perceptually motivated excitation pattern per time interval, that approximates the spectral resolution of the human auditory system. In the frequency domain, all energy that falls within the same ERB is summed up, while in the time domain a square-root Hanning window of length 0.0232 s with a 50% overlap was applied.

The ERB-representations of target and interferer are then compared with each other to determine which of the two sources dominates each individual time-frequency unit. From this comparison, two complementary binary masks for target and interferer are derived, which indicate at which regions the specified source is either dominant or inferior. The binary masks are then applied to the spectrograms of the mono signals. Afterwards, the resulting masked spectrograms are added to form a speech mixture which contains fragments of the locally dominant signal only. To impose spatial properties on the signal, head related transfer functions (HRTFs) for target and masker position are copied in the time direction and processed with the same binary masks as the mono signal. The resulting composite HRTF contains spatial properties of the dominant source in each time-frequency unit. After multiplication in the frequency domain, with the pre-processed mixture of mono signals and transformation into the time-domain, each signal fragment is moved to the desired location of either target or masker source.

The unaltered stimuli, which are used as reference, feature full HRTFs and full mono signals for both speakers. These stimuli contain simultaneous binaural cue differences and are referred to as HRTF stimuli.

Experimental Methods

In the experiment, speech reception thresholds (SRTs) for 50% intelligibility were measured using an alternative forced choice method (AFC). For the adaptive one-up-one-down procedure, the subjects were presented speech in noise or interfering speech via headphones. The speech material was taken from the OLdenburg LOgatome Corpus (OLLO) [5] and consisted of vowel-consonant-vowel (VCV) combinations with voiced middle phonemes. A typical trial is illustrated in the upper row of Figure 2. The target speaker was indicated by the signal word "ollo" prior to a sequence of six logatomes, that contained one differing logatome in one of the last three intervals.

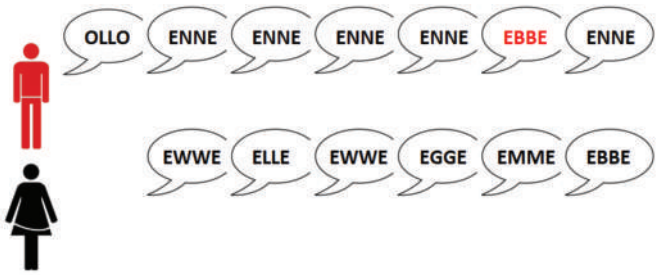

Figure 2: Illustration of a typical trial in the speech intelligibility measurement.

The listeners were instructed to attend to the target speaker and identify the deviating logatome out of a given list. Answers could be entered via a graphical user interface that also provided feedback about the correctness of the given answer. Nine paid listeners (four male, five female) participated in the experiment who were all native German speakers and normal hearing according to audiometric assessment.

Conditions

Two different masking conditions were chosen, varying in the amount of glimpsing opportunities. Speech being strongly modulated in time and frequency grants a large amount of glimpsing while stationary speech shaped noise allows only a minimum of glimpsing opportunities. In the speech masker condition the target speech was masked by simultaneously spoken logatomes from the same vowel subset. As indicated in the lower row of Figure 2, the interfering speaker was of opposite gender. In the condition with stationary speech shaped noise the masking noise spectrum was shaped according to the long term average spectrum of the target speaker. Throughout one

AFC run, the gender of the target speaker was kept fixed in order to account for possible deviations of the SRT due to voice characteristic of the target speaker.

In the condition with spatially separated sources, one of the two signals originated from an angle of 15° in the azimuth plane, while the other source was positioned at an angle of −15° creating a spatial separation of 30° between target and masker. For the collocated condition both sources were placed at either 15° or −15°. SRM was derived from the difference of $\mathrm{SRT}_{\mathrm{collocated}}$ and $\mathrm{SRT}_{\mathrm{separated}}$. For each condition (combination of spatial condition, masker type, speaker gender and manipulation) three repeated SRT measurements were conducted. To reduce influences due to training effects the first measurement was not included in the analysis.

Useful Speech Percentage (USP)

In order to gain more information about the amount of glimpses that were available in the presented speech mixture, a detailed analysis of the played back signals was performed. The aim was to determine the percentage of spectro-temporal units with a favorable local SNR. This measure is referred to as Useful Speech Percentage (USP). Other than the global SNR that represents an average over the entire trial (SRTs are displayed in global SNR), the USP allows a more detailed insight on masking in the time-frequency plane and accounts for the spectro-temporal interaction of target and masker. The analyzed speech material consisted of the reconstructed stimuli that had been used throughout the measurement phase of the AFC in the regarded condition. First, the spectro-temporal energy of target and masker was compared and the local SNR for each time-frequency unit was calculated according equation (1).

$$SNR_{i,j} = 10 \cdot log_{10}\left(\frac{PSD_{target,i,j}}{PSD_{masker,i,j}}\right) \qquad (1)$$

To distinguish the local SNR for one time-frequency unit from the global SNR that is being calculated over the length of the entire trial, the local SNR is referred to as $\mathrm{SNR}_{i,j}$. The indices i and j indicate the frequency-band number and number of time-frame. After excluding the silent time frames, leaving only time segments in which target speech was present, a histogram was computed that displays the distribution of local $\mathrm{SNR}_{i,j}$s during the trial. After setting a criterion, e.g. at -3 dB, the percentage of $\mathrm{SNR}_{i,j}$s that exceeded this criterion was determined. According to Cooke [4], a suitable criterion for considering a spectro-temporal unit as a glimpse, lies within a range of -2 to -5 dB. In the following analysis a criterion of -3 dB was applied. To obtain the final USP, the higher USP value out of the two ear channels was chosen, assuming that this was the better ear throughout the trial.

Results

Figure 3 shows the median speech reception thresholds in dB for speech in stationary speech shaped noise. SRTs obtained for unaltered stimuli (HRTF stimuli) that contained simultaneous binaural cues are shown on the left. SRTs for the manipulated stimuli (ISE stimuli) are displayed on the right.

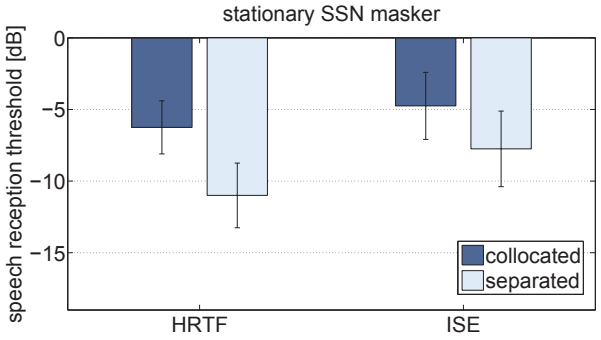

Figure 3: Median SRTs for the stationary-noise-masker condition. The SRTs for unaltered stimuli, containing simultaneous binaural cues, are displayed on the left. SRTs measured for the stimuli manipulated with ISE are displayed on the right. The errorbars display the standard deviations.

As can be seen in Figure 3 there is an overall raise in threshold for the ISE stimuli as compared to the HRTF stimuli. However, there's also a significant masking release for both stimulus types. The significance of differences between individual thresholds for the collocated and separated condition was assessed with a paired t-test, applying a 5% significance level. Thus, even without the availability of simultaneous interaural cue differences, a significant SRM can be achieved. Testing the differences between individual SRMs in the HRTF and in the ISE condition yields a significantly larger SRM for HRTF stimuli. The USP analysis reveals the proportion of glimpses that is available after the manipulation. For the noise masker condition the results of the USP-analysis is presented in Figure 4. Each bar shows the percentage of useful speech (amount of glimpses at the better ear) that were necessary to achieve a 50%-SRT.

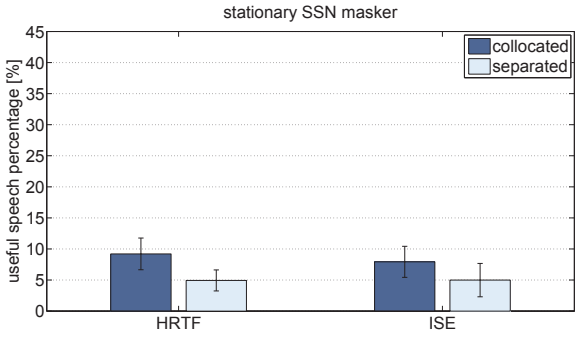

Figure 4: USP analysis for the stationary-noise-masker condition.

The USP analysis shows that for both stimulus types (HRTF and ISE) about an equal amount of glimpses is necessary to achieve the 50%-SRT in the separated condition. In the collocated condition the amount of necessary glimpses is not significantly different for the manipulated and reference stimuli.

SRTs for the speech-in-speech condition are given in Figure 5. In the case of a speech masker that allows a large amount of glimpsing opportunities, the median SRTs for both manipulated and unaltered stimuli were in the same range (-10 to -14 dB).

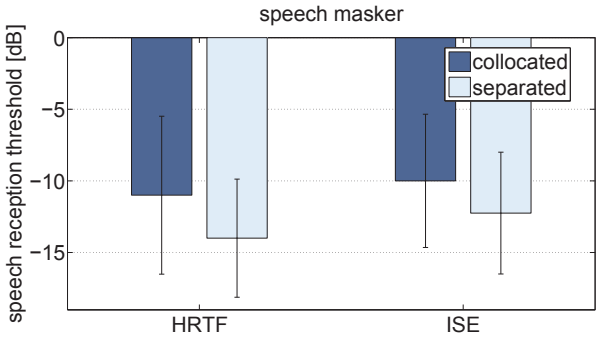

Figure 5: Median SRTs for the speech-masker condition.

In comparison with the speech-in-noise condition, thresholds in this condition are generally lower and the masking release between collocated and separated condition is less pronounced. Furthermore, the inter-individual differences are remarkably larger for the speech-in-speech task than for the speech-in-noise task. The averaged SRM for HRTF stimuli reaches significance on the 5%-level, whereas the averaged SRM for ISE stimuli does not reach significance. Both are, however, close to the significance criterion, suggesting that listeners' performance was not markedly different with and without available simultaneous binaural cues. When testing the listeners' individual SRM for both stimulus types, no significant difference was found between the SRMs for HRTF and ISE stimuli.

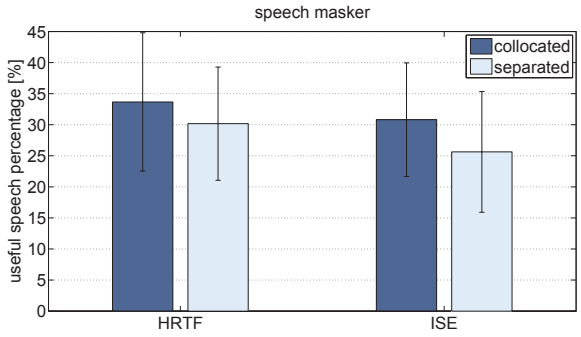

Figure 6: USP analysis for the speech-masker condition.

The USP-analysis in Figure 6 shows, that a somewhat smaller amount of glimpses was available after ISE manipulation, when the same performance (50% correctly identified logatomes) was achieved. Despite a lower USP in the ISE condition, the listeners' performance was still close to performance in the HRTF condition as can be seen from Figure 5. In comparison to the USP values of the speech-in-noise task, the values are higher in the speech-in-speech condition. This indicates, that in the presence of a noise masker, about 20% to 30% less useful speech is necessary to yield the same performance as in the presence of a speech masker. The remarkable threshold differences give an insight on the quantity of informational masking occurring in a multiple-speaker scenario.

Conclusions

An experiment was conducted to determine the spatial release from masking in the case of available simultaneous binaural cue differences and in the absence of such cues. Regarding the original question of how large the intelligibility benefit due to the availability of these cues is, it can be said that in the case of a speech-in-speech task no significant difference in SRM was found between a condition with these cues available and the condition of their absence. This outcome indicates that the information eliminated after the ISE manipulation, such as simultaneous binaural cue differences and the presence of target signal in spectro-temporal regions with low SNRs, is somewhat redundant for SRM. Consequently, the BMLD-like mechanism does not seem to contribute significantly to speech intelligibility in the investigated scenario.

In the presence of a stationary speech shaped noise masker, the spatial release from masking differed significantly between the manipulated and the unaltered stimuli. However, the USP-analysis reveals that the amount of spectro-temporal regions with favorable $\mathrm{SNR}_{i,j}$ was similar in both conditions, suggesting that the listeners' performance is explicable with the amount of available glimpses.

Results of the USP-analysis suggest that drops in performance after the ISE-manipulation can likely be attributed to the reduced amount of glimpses. This suggests that advantages in speech intelligibility due to spatial separation strongly relies on the information drawn from glimpses at the better ear. Assuming a criterion value of -3dB, the USP measure seems to be a suitable predictor for speech intelligibility.

References

[1] Moore, B.C.J.: An Introduction to the Psychology of Hearing. Academic Press, 1997.

[2] Durlach, N.I.: Equalization and Cancellation Theory of Binaural Masking-Level Differences. J.Acoust.Soc.Am. 35(1963), 1206-1218.

[3] Beutelmann, R. and Brand, T.: Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. J.Acoust.Soc.Am. 120(2006), 331-342.

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

[5] Wesker, T., Meyer, B., Wagener, K., Anemüller, J., Mertins, A. and Kollmeier, B.: Oldenburg Logatome Speech Corpus (OLLO) for Speech Recognition Experiments with Humans and Machines, in: Proceedings of Interspeech, Lisbon, Portugal, (2005), 1273–1276.