Buenos Aires – 5 to 9 September, 2016
Acoustics for the 21st Century…

PROCEEDINGS of the 22nd International Congress on Acoustics

Communication Acoustics: Paper ICA2016-496

On the contribution of spatial hearing to speech intelligibility

Steven van de Par[(a)], Sarinah Sutojo[(a)], Esther Schoenmaker[(a)]

(a) Acoustics Group, Cluster of Excellence "Hearing4all", Carl-von-Ossietzky Universität Oldenburg, Germany, {Steven.van.de.Par, Sarinah.Sutojo, Esther.Schoenmaker}@uni-oldenburg.de

Abstract

A spatial separation between a target speaker and interfering sources is known to improve speech intelligibility. Various effects may contribute to this: the target-speech-to-interferer ratio may be better in one ear, binaural unmasking may improve the detectability of target speech components, and spatial cues may facilitate better segregation of the target speech from the background. An experiment will be discussed that uses a stimulus manipulation that reduces the possibility of having a binaural unmasking effect. Interestingly, these manipulations have little effect on speech intelligibility and this small effect can be explained based on the percentage of glimpses that is available in the stimulus. In a second experiment, the stimuli are manipulated in such a way that only the most salient proportion of the target speech glimpses is located at a different position than the interfering speech sources. This very small amount of spatial cues is already enough to create a strong spatial advantage in speech intelligibility. The presence of the sparse spatial cues seems to create a more effective processing of monaural cues in line with ideas of auditory grouping and stream segregation.

Keywords: Spatial Hearing, Speech Intelligibility, Binaural, Stream Segregation

On the contribution of spatial hearing to speech intelligibility

1 Introduction

The human auditory system has a remarkable capacity to selectively understand one specific speaker in the presence of interfering speech [1]. Speech intelligibility is possible even when the target speaker is lower in level than the interfering sound sources. In part, good speech intelligibility can be achieved due to the fact that speech is a spectro-temporally sparse signal. As a consequence, when various speech sources are present at the same time, due to the fluctuating levels of all speakers, the target speaker can be locally higher in level than the interfering speakers [1]. Therefore, it can be expected that the total amount of energetic masking, which would occur when the target speaker is locally lower in level than the interferers, is relatively small. In line with the significance of spectro-temporal sparseness, it has been found that speech intelligibility in noise improves when the noise is fluctuating over time [2].

Cooke [3] argued that listeners could use the portions of the target speaker that are not energetically masked in order to understand speech. These so-called glimpses are in principle available to achieve a good degree of speech intelligibility. However, when multiple competing speakers are present, it is a challenge for the auditory system to know which glimpses are uttered by the target speaker. The general properties of the target and interfering speakers can be expected to be fairly similar and distinct cues such as differences in pitch range, voice timbre, and spatial position of the sources are needed to differentiate between signal components that belong to the target and to interfering speakers.

In many experiments it has been demonstrated that a difference in the spatial position of the target and interfering speakers leads to an improved speech intelligibility as compared to spatially co-located sources [4]. This may be in line with a better selection of the target signal components across the spectro-temporal plane, based on spatial cues. [5].

An alternative explanation, however, would utilize the reduced level of energetic masking that results from spatially separated masker and target components [6,7]. This effect has inspired various models for predicting speech intelligibility which implement an Equalization Cancellation stage that reduces the internal level of interfering speakers when target and interferers are spatially separated [8,9].

In this contribution two sets of experimental conditions will be discussed that investigated the role of spatial cues in improving speech intelligibility, followed by a discussion that focusses on the way in which spatial cues facilitate speech intelligibility. The first experiment was part of a previous study [10] and will be used as a comparison condition for the second experiment. Both experimental conditions present speech intelligibility data for a target speaker in the presence of two interfering speakers. In the first experiment the target speech is manipulated in such a way that target speech components are removed that have a local signal-to-noise ratio that is lower than a certain threshold SNR criterion measured within a small spectro-temporal interval. By removing these low-level target components it can be investigated at what SNR criterion, target

components start to contribute to speech intelligibility. If binaural unmasking improves speech intelligibility, low-SNR target components should contribute more to speech intelligibility in spatially separated speaker conditions than in collocated conditions.

In the second experimental condition, only the highest level target components are spatially separated from the interfering speakers as determined on a spectro-temporal basis. This should give insight in how much spatial separation is needed in the stimulus in order to elicit a benefit of spatial separation.

2 Experiments

The goal of these experiments was to get insight into the importance of spatial cues in speech intelligibility. The focus is twofold. Firstly, to gain insight into the contribution of spatial unmasking, it was investigated whether target speech components with a low spectro-temporal SNR contribute differently for collocated and separated speakers (cf. [10]). If binaural unmasking contributes to improving speech intelligibility, components with a lower spectro-temporal SNR should have more importance in case of separated speakers. A second question was whether spatial cues contribute to speech intelligibility by allowing a better stream segregation. This was investigated by applying informative spatial cues only to the most salient (highest spectro-temporal SNR) components of the target speaker that should not be subject to any energetic masking.

2.1 Stimuli

Target speech sentences were selected from the German OLSA speech corpus which consists of unpredictable five-word sentences where each word can be one out of 10 alternatives [11]. The structure of such a sentence was “name – verb – numeral – adjective – object”. The target sentences spoken by a male speaker were presented simultaneously with two interfering speech utterances, produced by different female speakers, which were random selections from two audio books. All three speech sources were either convolved with the same zero-degree azimuth Head Related Impulse Responses such that they were spatially collocated, or with three different HRIRs at azimuths of -60, 0, and +60 degrees for the spatially separated condition; the target speaker was presented at 0 degrees. The target was presented at -11 dB relative to the total level of both interfering speakers, measured after HRIRs manipulations were applied. All stimuli were presented over Sennheiser HD 650 earphones at a level of 65 dB SPL within a sound proof booth.

In order to investigate the role of binaural unmasking in speech intelligibility, first the local SNR was determined for each 1-ERB wide, 23-ms long unit of the spectro-temporal plane. The local SNR was determined by comparing the target speaker to the sum of the two interfering speakers after spatialization. When the local SNR was lower than a predetermined criterion SNR value, the corresponding spectro-temporal unit of the target speaker was set to zero using a Fast-Fourier-Transform based overlap-add filtering, in this manner erasing the low local SNR parts of the target speaker entirely.

In the second condition, in order to investigate the role of spatial cue-based segregation, the local SNRs were determined across the spectro-temporal plane in the same manner as previously described. Only now, if the local SNR was below a certain SNR criterion value, the corresponding spectro-temporal unit of the target speaker was spatially rendered to the location of the simultaneously most dominant interferer (-60 or +60 degrees). All other target units were rendered at 0 degrees, being at a different location than the interfering speakers. Note that in this condition, all spectro-temporal units of the target speaker were present in the stimulus, only their spatial positions were manipulated.

2.2 Procedure

A total of 6 normal-hearing listeners were presented with five-word target sentences in the presence of two interfering speakers. The task of the listeners was to select the five words in a user interface that allowed for 10 alternatives for each of the five words. In this OLSA matrix test, the 10 different alternatives for each word were part of a closed set that stayed the same across the full experiment [11]. Subjects were trained to familiarize themselves with the speech material. Each unique stimulus condition was tested with 20 different sentences. The number of correctly reported words was used to determine the percentage correct.

The following conditions were tested in the experiment. For the unmasking condition, both collocated and separated spatial conditions were tested. The criterion SNR value was varied in several steps to assess for what SNR values target speech components start to contribute to speech intelligibility. For the segregation condition, only spatially separated conditions were presented also with various criterion SNR values, to assess whether it is sufficient that only the most salient target speech components are spatially separated from the interfering speakers.

2.3 Results

The results for the first condition, testing for the role of binaural unmasking, is shown in Fig. 1. As expected, the percentage of correct word recognition is higher for spatially separated as compared to spatially collocated sentences. For the collocated condition (open symbols), it can be seen that components below a local SNR of -5 dB do not contribute to speech intelligibility. When removing target speech components above this level, a gradual reduction in speech intelligibility is seen.

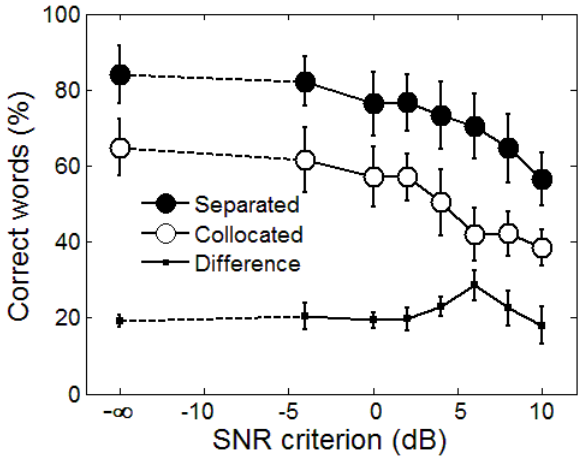

Figure 1: Percentages correct word recognition as a function of the SNR criterion. All speech components with a local SNR below the SNR criterion were removed (cf. [10]).

Based on binaural unmasking, it could be expected that target speech components that are energetically masked in the collocated condition, would not be masked in all cases for separated conditions, and thus are in principle available to contribute to speech intelligibility. If these low local SNR components did indeed contribute, one would expect to see a degradation in performance already at lower local SNR criterion values for the separated condition. The data for the separated speakers (closed symbols) do not show this effect. Instead the pattern of data seems to be rather parallel to the collocated data pattern (as indicated by the difference between both, indicated by the point-marked curve).

Also of interest is the improvement that is seen above SNR criterion values of -4 dB for the spatially separated condition as compared to the collocated condition. In this case one could expect that all the target speech components should be audible; i.e. energetic masking is not a factor in reducing speech intelligibility. Even if all target speech components are audible, spatial cues still help to improve speech intelligibility. This would be in line with the assumption that spatial cues help to resolve what stimulus components belong to the target speaker.

Fig. 2 shows the results for the condition where only the target components above a certain criterion SNR value are spatially separated from the interfering speakers (upper black curve). The curve for removed target components for the separated condition that was already shown in Fig. 1 is replotted in Fig. 2 (lower grey curve). It can be seen, that for low SNR criterion values, both curves are essentially the same; the small differences that are seen, are probably due to differences in training across the two conditions.

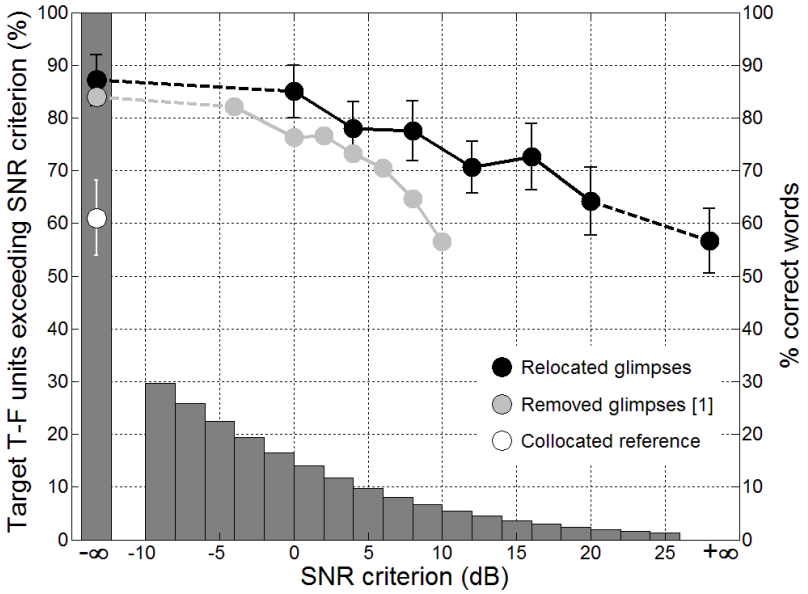

Figure 2: Percentages correct word recognition as a function of SNR criterion. The upper (black) curve shows results for target components (glimpses) being relocated to interferer positions when below the SNR criterion. The lower (grey) curve shows results of Figure 1, where target components (glimpses) are removed below the SNR criterion. The white data point is the collocated reference condition without any manipulation. The bar plots show the percentage of target components (TF-units) exceeding the SNR criterion.

It can be seen that the condition with relocated glimpses shows a much slower decrease towards increasing SNR criterion values than the condition with removed glimpses. The fact that we found a difference between both conditions in the range of criterion SNRs between 5 and 10 dB demonstrates that the glimpses below the criterion value, which are not spatially separated from the interferers, still contribute to speech intelligibility. Especially, at the criterion SNR of 10 dB there is a difference of about 15% correct word recognition which is caused by the presence of target speech glimpses that are not separated from the interferer.

Note also that for the criterion SNR value of 10 dB, only about 6 percent of the target glimpses are spatially separated from the interferers. This small percentage is sufficient to gain more than 15% in correct word recognition as compared to all target glimpses being collocated with the interferers (corresponding to the rightmost point in Fig.2). This is more than half of the total improvement that can be obtained when all target glimpses are spatially separated from the interferers.

3 Discussion

The experimental results that are discussed above indicate that binaural unmasking phenomena similar to the reduction in energetic masking that is seen in 'classical' Binaural Masking Level Difference experiments [6,7] is not a significant factor explaining the improved speech intelligibility that is seen for spatially separated as compared to collocated speakers. In contrast, results seem to indicate that despite the fact that target speech components are clearly above the energetic masking threshold, the presence of spatial cues still improves speech intelligibility. Such an improvement is in line with spatial cues contributing to a better selection of target speech components from the complex stimulus; i.e. an improved auditory stream segregation.

The results also show that if only a small portion of target glimpses is spatially separated from the interfering speech sources, the remaining, collocated glimpses still contribute to improved speech intelligibility. When only the 8% most salient target components are spatially separated from the interfering speakers, this already amounts to about half the improvement that can be obtained with a fully spatially separated target source. This may indicate that salient, spatially separated glimpses can serve as an anchor for speech elements that have already been grouped monaurally.

It is an interesting question why the binaural unmasking effect which has been demonstrated in various basic experiments with tones in noise does not seem to contribute in this more complex speech-in-speech conditions. A possible limiting factor may be that, even if target speech components are retrieved due to binaural unmasking, the auditory system will not be able to infer whether the retrieved component belongs to the target or to one of the interfering speakers. Specifically this may be so because target speech localization cues are not well represented after binaural unmasking.

The conditions that have been discussed here present the significant challenge for the auditory system to select stimulus components that belong to the target speaker. There are two primary cues that can be used; the target speaker is a male speaker, while the interfering speakers are female, and the target speaker is in central position which is a cue that is only useful in the separated condition.

The results of our experiments have shown that already with 8% of the most salient target glimpses being located at different locations than the interferers, one can observe a considerable improvement in speech intelligibility. In such a mix of three speakers as used in our experiments there will be various spectro-temporal regions where the target and interferer are similar in level. In this case, it can be expected that the spatial cues of the target speaker will be distorted by the co-presence of the interfering speakers that have different spatial cues. Considering this, it would be a good strategy for auditory processing to put more weight on spatial cues that have a high local SNR. The auditory system may infer the high local SNR from the local coherence of the signal.

The use of spatial cues in the rapidly fluctuating signal, where dominance of the target speaker quickly interchanges with a dominance of one of the interfering speakers, requires the auditory system to be able to extract spatial cues from the auditory complex stimulus with a high

temporal resolution. In addition, the binaural system needs to infer the reliability (i.e. by extracting the interaural coherence) based on brief glimpses. Such assumptions are part of some current binaural modelling approaches such as e.g. [12].

Acknowledgments

This work was supported by the DFG funded Collaborative Research Center, Transregio 31: The active auditory system.

References

[1] Brungart, D.S.; Chang, P.S.; Simpson B.D.; Wang D. Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *J. Acoust. Soc. Am.*, Vol. 120, 2006, pp. 4007-4018.

[2] Festen, J.M.; Plomp, R. Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *J. Acoust. Soc. Am.*, Vol. 88, 1990, pp. 1725-1736.

[3] Cooke, M. A glimpsing model of speech perception in noise. *J. Acoust. Soc. Am.*, Vol. 119, 2006, pp. 1562-1573.

[4] Bronkhorst, A. W. The cocktail party phenomenon: A review of research on speech intelligibility in multi-talker conditions.*Acoustica*, Vol. 86, 2000, pp. 117-128.

[5] Schoenmaker, E.; Brand, T.; van de Par, S. The multiple contribution of interaural differences to improved speech intelligibility in multitalker scenarios. *J. Acoust. Soc. Am.*, Vol. 139, 2016, pp. 2589-2603.

[6] Zurek, P.M.; Durlach, N.I. Masker-bandwidth dependence in homophasic and antiphasic tone detection.. *J. Acoust. Soc. Am.*, Vol. 81, 1987, pp. 459-464.

[7] Van de Par, S.; Kohlrausch, A. Dependence of binaural masking level differences on centre frequency, masker bandwidth, and interaural parameters. *J. Acoust. Soc. Am.*, Vol. 106, 1999, pp. 1940-1947.

[8] Beutelman, R.; Brand, T. Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. *J. Acoust. Soc. Am.*, Vol. 120, 2006, pp. 331-342.

[9] Wan, R.; Durlach, N.I.; Colburn, H.S. Application of a short-time version of the equalization-cancellation model to speech intelligibility experiments with speech maskers. *J. Acoust. Soc. Am.*, Vol. 136, 2014, pp. 768-776.

[10] Schoenmaker, E.; van de Par S. Intelligibility for Binaural Speech with Discarded Low-SNR Speech Components. *Advances in experimental medicine and biology*, Vol. 894, pp. 73-81.

[11] Wagner, K.; Brand, T. Kollmeier, B. Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests. *Zeitschrift für Audiologie*, 38, 1999, pp. 20-24.

[12] Faller, C.; Merimaa, J. Source localization in complex listening situations: Selection of binaural cues based on interaural coherence, *J. Acoust. Soc. Am.*, 116, 2004, pp. 3075-3089