Speech intelligibility in realistic listening situations for different numbers, azimuths and movement of speech or noise maskers

Martin Hansen, Felix Dollack, Geske Linnéa Eberlei,
Hannah-Lina Grahlmann, Wiebke Lamping, Stefan Raufer
Institut für Hörtechnik und Audiologie, Jade Hochschule, 26121 Oldenburg, Germany, martin.hansen@jade-hs.de

Introduction

Speech intelligibility in a noisy environment can be measured with different standardized test methods. One very common experiment setup uses, e.g., a single loudspeaker for the presentation of speech and noise with frontal incidence in a S_0N_0 paradigm. Sometimes the SRT is additionally measured in a S_0N_{90} with two loudspeakers, in order to estimate the size of a binaural intelligibility level difference (BILD). The most commonly used masker signal is unmodulated, speech shaped noise. While these tests are well-described, easily reproducible and their resulting SRT are well-known, most of these speech audiometric methods lack a close correspondence to real-life situations. For example, sound signals in everyday life can arrive from different directions, rather than being presented by one or a few fixed loudspeakers.

Jensen[1] and Lamping and Hansen [2] investigated the influence of an uncertainty of the target location/azimuth on the intelligibility of speech in noise from fixed directions. In the current study, the opposite was tested: speech reception thresholds (SRTs) were measured for a target with a fixed frontal incidence, while different numbers of maskers were placed at random azimuth and distance around the subject. The sound sources were generated in a wave field synthesis (WFS) system which allowed a highly flexible placement of the different masker sound sources, including a possible movement of the masker sound sources around the test subject. The use of a WFS offers the advantage that the measurements could easily be repeated with hearing impaired subject using hearing aids.

Experimental setup

Stimuli

The target speech material consisted of the German matrix sentence test (“OLSA”) [3], containing five word sentences with all-identical grammatical structure as in “Peter kauft sieben nasse Sessel”. The level of the speech was adaptively varied to converge to the speech reception threshold (SRT), defined as the S/N for 50% words correct. The level of the masker(s) at the listening position was held constant at 65 dB SPL, regardless of the number of maskers presented (see below).

Two types of masker signals were used in the study, either noise or intelligible conversations. The noise masker was the standard noise accompanying the OLSA matrix test, i.e. a spectrally speech-shaped, temporally unmodulated noise. The conversation maskers, which consisted of intelligible speech, were generated from two-person conversations taken from several radio interviews. The two speakers of one conversation never talked at the same time. Rather, the conversation changed over from one speaker to the other, and each conversation was temporally trimmed to exhibit 1 speaker changeover during the duration of the target sentence. This is illustrated in Fig. 1. Any temporal gaps in the conversations were limited to a duration of max. 60 ms, in accordance with [1]. Each conversation masker was linearly filtered to yield a match to the long term average spectrum of the speech target stimuli.

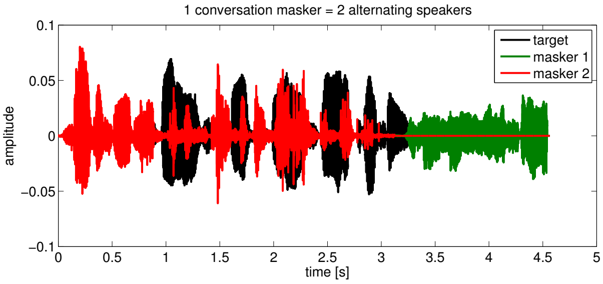

Figure 1: Temporal waveform of a conversation masker, exhibiting a changeover between the two speakers (red and green), in the presence of the target sentences (black). The conversation masker starts approx. 1 s prior to the target sentence.

WFS-rendering

All stimuli, targets and maskers, were generated as virtual point sound sources using a commercial WFS system. The room containing the WFS setup is shown in Fig. 2. The WFS loudspeakers are placed equidistantly, with Δx=6 cm for the tweeters, along four arrays forming a rectangular listening area with dimension 4 m × 5,5 m. The WFS-loudspeakers are placed at ear-height of the

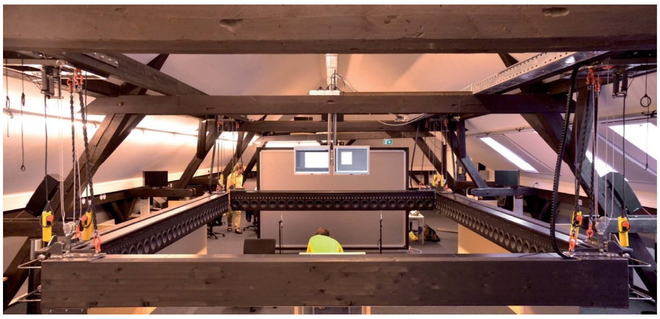

Figure 2: View of the WFS system in the “3D-Lab” at Jade Hochschule.

subject sitting on a chair in the centre of the rectangular WFS array.

The target sentence was always presented at the same position, at 1 m distance from the test subject with frontal incidence.

For both masker types, a number of n=1,2,3,4, or 8 uncorrelated maskers were presented at positions with random azimuth angles around the test subject in order to simulate a cocktail party situation. Each masker could have a random distance from the subject ranging between 1 m and 15 m. A new random position of the masker(s) was chosen for every new target sentence. To avoid incidentally coinciding positions of the maskers, every masker had a "privacy" sphere of 60 cm radius where no other masker could be placed. An example for the placement of 8 maskers, including their privacy spheres, in shown in Fig. 3. The masker(s) could either be fixed in place during one target sentence, or could move circularly around the subject at a speed of 3 m/s, while the target speech signal still always came from the front.

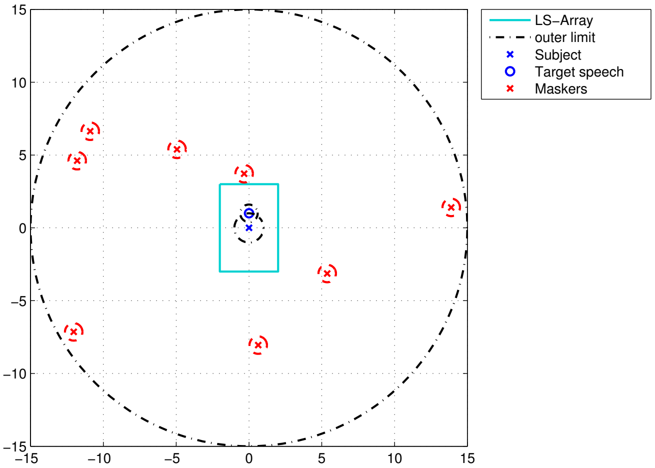

Figure 3: Example of the placement of 8 maskers around the test subject in the centre of the setup. The dimensions are given in Meter.

Procedure

22 normal-hearing subjects participated voluntarily in the study. The age of the subjects ranged between 22-44 years and males/females were balanced.

Lists of 20 sentences were used for each measurement. The first two sentence lists were used solely for training and the results were discarded for every subject. Subsequent to this training, each subject went through a randomized order, including one repetition each, of the following test conditions:

- $n = 1, 2, 3, 4, 8$ masker noises, random directions, *fixed*
- $n = 1, 2, 3, 4, 8$ masker conversations ($\equiv 2, 4, 6, 8, 16$ different speakers), random directions, *fixed*
- $n = 1, 8$ masker noises / masker conversations, random directions, *rotating around subject*

Results

The results from the 22 normal-hearing subjects are shown in Fig. 4.

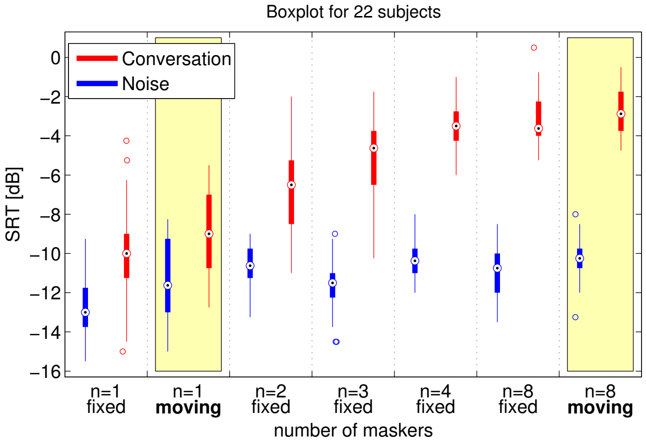

Figure 4: Boxplots of the SRTs of 22 subjects for the different test conditions.

Noise maskers

Looking at the noise maskers first, a statistical analysis showed that the n=1 fixed noise yielded a significantly different SRT compared to all other noise masker conditions, whereas mainly no significant differences were observed among all other noise masker conditions. This meant, amongst others, that the introduction of a rotation of the masker(s) around the subject *had* a significant effect for n=1, but *not* for n=8, relative to the same conditions with fixed masker(s).

Speech maskers

For the speech maskers, a statistical analysis showed a significant increase of SRT with increasing number of the fixed maskers for $1 \leq n \leq 4$. For $n \geq 4$, a saturation effect could be observed, meaning that a further increase of the number of maskers beyond $n \geq 4$ did not result in a further significant increase of the SRT. For both n=1 and n=8, the introduction of a rotation of the masker(s) showed *no* significant difference relative to the same situation with fixed masker(s).

Discussion

The SRT result for the n=1 fixed noise masker condition can be somewhat compared to the result for a standard S_0N_{90} loudspeaker setup, which is known to lead to a slightly lower SRT and a similar spread among subjects. In the S_0N_{90} condition, the interaural differences of the noise signal will be maximal and an accordingly strong spatial release from masking effect ("BILD") can be expected relative to the S_0N_0 reference condition with its SRT of -7,1 dB according to literature [3]. In the current experiment, however, the noise source can have any arbitrary azimuth. Some effect of spatial release from masking can therefore be expected for the majority of times, varying continuously with masker azimuth(s). On average, the more seldom random cases where the noise's azimuth was close to that of the target at 0°, i.e. no spatial masking release, will have a minor overall effect during the course of 20 sentences used for one SRT measurement. The difference between the presentation by either a real loudspeaker or a virtual WFS sound source is not expected to show any effect on speech intelligibility here. The varying radius of noise source is neither expected to yield any effect on SRT here, as the level of the noise was always fixed at the listeners position. In sum it is concluded that the randomness of the noise azimuth, which introduces some amount of uncertainty for the subject and therefore possibly an extra load on her working memory, plays a minor role on the SRT compared to fact that the noise signal is spatially separated from the target signal.

Going from $n = 1$ to $n \geq 2$ for fixed noises led to an increase of the SRT by 2.2 dB on average, with an unexplained small deviation for $n = 3$. This observation is in line with the idea of a purely energetic masking effect for two or more noise maskers, as the masker level was always kept constant, regardless of n.

On the contrary, a comparison with the SRT data for the conversation maskers showed an increase of SRT with increasing masker number for $1 \leq n \leq 4$. This marked extra masking property of the conversation maskers relative to the noise maskers indicates a strong effect of informational masking. This effect can be explained by the intelligibility of the conversation maskers which puts a high extra cognitive load on the listener. The fact that the target and the conversation maskers are spatially separated might “help” the subject to differentiate between the auditory streams of target and masker(s). However it might also increase the intelligibility of the maskers further, compared to a S_0N_0 situation, thereby increasing their distracting effect on the subject. This is in line with the observation that an extra conversation masker, being again spatially separated from both target and also the other masker(s), *decreased* the overall intelligibility of the target, at least up to $n \leq 4$.

The use of temporally modulated or speech-like maskers often promotes the so-called dip-listening effect, which means that pauses or soft parts in the masker lead to an increased intelligibility of the corresponding portions of the target compared to an unmodulated masker at the same level. Although other temporal portions of the target are then masked more effectively, the net effect is an increased intelligibility. The corresponding SRT can sometimes be found to decrease by more than 10 dB. Such an effect was clearly not observed in the current study. The reason is most likely that the duration of pauses in the conversation masker was limited to max. 60 ms which is clearly shorter than in naturally occurring speech.

References

[1] Niels Søgaard Jensen, René Burmand Johannsesson, Søren Laugesen, and Renskje K. Hietkamp. Measuring speech-in-speech intelligibility with target location uncertainty. In *ISAAR*, Nyborg, 2011.

[2] Wiebke Lamping and Martin Hansen. Messung der Sprachverständlichkeit bei veränderlicher Nutzsignalrichtung in einem Wellenfeldsynthese-system. *Z. Audiol.*, 51(4):134–139, 2012.

[3] Kirsten Wagener, Volker Kühnel, and Birger Kollmeier. Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests. *Z. Audiol.*, 38(1):4–15, 1999.