The influence of dynamic binaural cues on speech intelligibility at low and high frequencies

Jan Heeren[1], Giso Grimm[2], Volker Hohmann[3]
[1,2,3] *Universität Oldenburg and Cluster of Excellence Hearing4all, 26111 Oldenburg*
E-Mail: [1] *j.heeren@uni-oldenburg.de*

Abstract

The amount of spatial release from masking is mainly determined by the change in interaural time difference (ITD) of the noise relative to the ITD of the signal [12]. Accordingly, speech-in-noise with frontal speech presentation and noise from the front or back (S0N0 and S0N180 conditions) lead to similar detection and speech intelligibility thresholds. However, head movements can introduce dynamic binaural cues that may lead to a release from masking (RFM). In this study the effect of dynamic binaural cues on speech intelligibility was investigated for lowpass and highpass filtered signals to assess the influence of ITDs and interaural level differences. Movements were implemented as modulations of the nominal azimuths of the sound sources (S0N180, S0N0). These modulations were either in-phase or anti-phase for S and N. The stimuli were rendered using eleventh order ambisonics with 'basic' decoding, and presented via loudspeakers. Speech and noise signals were filtered at 1000 Hz (lowpass), 1500 - Hz (highpass) or unfiltered. Results show a significant RFM with dynamic binaural cues for S0N0 in all filter conditions. For S0N180 only the unfiltered condition shows a significant RFM.

Funded by DFG FOR1732

Introduction

The intelligibility level difference (ILD) is a measure for the spatial RFM of speech-in-noise. It is defined as the difference in speech reception thresholds (SRT) between conditions with collocated speech and noise in the front (S0N0) and conditions with speech in the front and noise from various azimuths (S0NX):

$$\mathrm{ILD}(X) = SRT_{S0N0} - SRT_{S0NX} \quad \text{[dB]} \qquad (1)$$

ILDs are minimal at S0N180 (0 - 3 dB) and show high values if the noise is presented from the side (maximum 13 dB at 120°) [3]. Thus, the speech signal is masked most effectively if speech and noise are located on a front-back axis in the median plane. For condition S0N180 and the reference condition S0N0 the reason seems to be obvious, because these conditions are characterized by a lack of binaural cues. The only cue that influences condition S0N180 compared with S0N0 is the pinna effect, which affects rear signals by damping high frequencies. “Front-back masking” also occurs off the median plane, though. Results from Saberi et al. [12] show that signals are generally masked most effectively if the noise is located on the same front-back axis, e.g., the combinations S30N30 and S30N150 lead to similar thresholds. This was shown for click trains in white noise.

Ambiguities between the front and the rear hemisphere could be related to the fact that ITDs are equal for the front and the back hemisphere. A known example for an effect of ITD ambiguities is the occurrence of front-back-confusions in localization experiments. Evidence for this relation was for example provided by Brungart and Simpson [4], who found that the rate of front-back confusions is depending on the presence of high frequencies. For lowpass filtered noise signals below 1 kHz the rate of front-back confusions was around 40% (chance level: 50%), whereas unfiltered signals only showed 5% front-back confusions. ITDs are only relevant for frequencies up to around 1 kHz while being the dominant cue for localization in this frequency range [14]. These findings support the assumption for a causal relation between the occurrence of front-back confusions and the ambiguity of the ITDs. At high frequencies the dominant cues for both localization and masking are interaural level differences.

Front-back confusions can be resolved by head movements that introduce dynamic binaural cues [1]. Heeren et al. [10] adapted this approach and showed that dynamic binaural cues, which were introduced by movements of virtual sound sources, can lead to a significant RFM in front-back masking situations. The method was based on a separate assessment of detrimental and beneficial movements that were characterized by the relation of ITDs between signal and noise. The question remained whether the observed SRTs are generally related to the ambiguity of ITDs and if the occurrence of front-back confusions is critical for speech intelligibility. Therefore, the experiment was repeated here with lowpass and highpass filtered speech and noise to investigate whether the increased occurrence of front-back confusions for lowpass filtered signals and the amount of dynamic RFM correlate.

Method

Movement conditions

Measurements were based on the approach introduced by Heeren et al. [10]. Conditions S0N0 and S0N180 were tested for three movement conditions:

Rot) S and N moving rotationally

Cou) S and N moving counter-rotationally

Stat) Static reference condition

The movements were implemented as modulations of the nominal azimuths of S and N (1 Hz, 10°), resulting in oscillations of the virtual sound sources around their nominal positions. To create the rotational and the counter-rotational relation, modulations were either in-phase for S and N or anti-phasic. Figure 1 shows a sketch of the movement conditions for S0N180. A comparison of these conditions reveals the influence of spatial RFM and the pure influence of movements separately. On the one hand movements with constantly equal ITDs for S and N do not lead to a RFM and represent the pure movement effect (S0N0Rot, S0N180Cou). These movements are expected to be detrimental for speech intelligibility and results will be displayed in red. On the other hand movement conditions with opposed ITDs lead to a RFM and are expected to be beneficial for speech intelligibility (S0N0Cou, S0N180Rot). Results for these conditions will be colored green. These "green SRTs" are influenced by two factors: the spatial RFM that is caused by temporal lateral displacement of the sound sources, and the pure movement effect. By subtracting "red SRTs from "green SRTs" the pure movement effect can be excluded and a dynamic spatial RFM component can be derived.

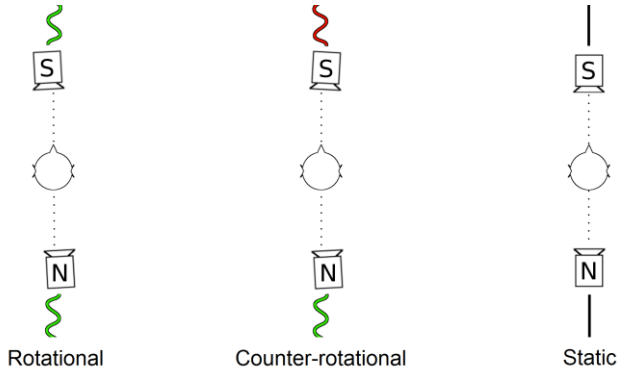

Figure 1: Sketch of the three movement conditions for S0N180; source movements were implemented as a modulation of the nominal azimuths; the modulations were either in-phase or anti-phasic for S and N resulting in rotational or counter-rotational sound source movements; the third condition was static.

Speech-in-noise test

SRTs were measured using a german matrix test called Oldenburg Sentence Test (OLSA) [13]. It consists of sentences with a fixed structure name-verb-numeral-adjective-object. These were presented against the corresponding stationary speech-shaped noise (OlNoise), that was presented at a fixed level of 65 dB SPL while the speech level was adjusted adaptively towards the 50%-speech reception threshold. One measurement list contained 20 sentences. The three movement conditions (Rot, Cou, Stat) were tested in interleaved order. Thus, a measurement run consisted of 3x20 sentences.

Additional to the original speech test, the measurements were conducted using lowpass and highpass filtered speech and noise. This was realized applying fifth-order Butterworth filters at the cutoff frequencies of 1 kHz (lowpass condition) and 1.5 kHz (highpass condition). Filter conditions were tested in randomized order. A training list was performed for each filter condition, followed by the spatial conditions S0N0 and S0N180 in randomized order.

Setup

Stimuli were presented using a horizontal loudspeaker array consisting of 24 loudspeakers (Genelec 8020). It was the same setup as in the Heeren et al. [10] study. Loudspeakers were set up regularly on a circle with a radius of 2 m. It was placed in the Communication Acoustics Simulator in the House of Hearing, Oldenburg, which is a sound treated room with a reverberation time of 0.4 - 0.6 s (T60) [2]. Loudspeakers were equalized and phase delays were compensated. This was realized by a division of the output signals for each loudspeaker by its impulse response (IRS) in the frequency domain. The IRSs were recorded previously using a Neumann KM183 microphone that was placed in the center of the loudspeaker setup. To use the direct sound part of the IRS only, they were shortened to a length of 2.4 ms, which was determined by the delay between the IRS onset and the first reflection. Thus, the compensation does not affect frequencies below 400 Hz. Virtual sound sources were panned using an eleventh order basic ambisonics algorithm [11] that is part of the Toolbox for Acoustic Scene Creation and Rendering (TASCAR) [6,7,8]. During the measurements participants were placed on a chair in the center of the loudspeaker setup. They were instructed to look at a fixed point at 0° azimuth and keep their heads still. The perceptual spatial resolution of participants in this setup was evaluated by measuring minimum audible angles with the Olnoise stimulus. A median MAA of 2.7° was measured for a reference azimuth of 0°, which is between literature values for white noise in anechoic rooms and reverberant rooms [9].

Participants

Twelve normal hearing subjects participated in the experiment (seven male, five female, age: 21-42 years). All of them showed hearing thresholds of below 20 dB HL (0.1-8 kHz) and had prior experience with speech-in-noise tests.

Results

Figure 2 shows a boxplot of the measured SRTs for S0N0 and S0N180 with filtered and original stimuli and the three movement conditions (medians and interquartile ranges). The results of the original condition are taken from the earlier study [10]. In the static reference condition median SRTs of -7.8 dB (S0N0) and -9.8 dB (S0N180) were observed with unfiltered signals. Lowpass filtering led to shifted values of -4.8 dB (S0N0) and -6.2 dB (S0N180), while the highpass condition shows values of -6.0 dB (S0N0) and -7.8 dB (S0N180) for static listening conditions.

For the analysis of the relations between the movement conditions, the medians and interquartile ranges of the dynamic RFM are displayed in figure 3, which are calculated by the individual differences between the red SRT and the green SRT per spatial condition.

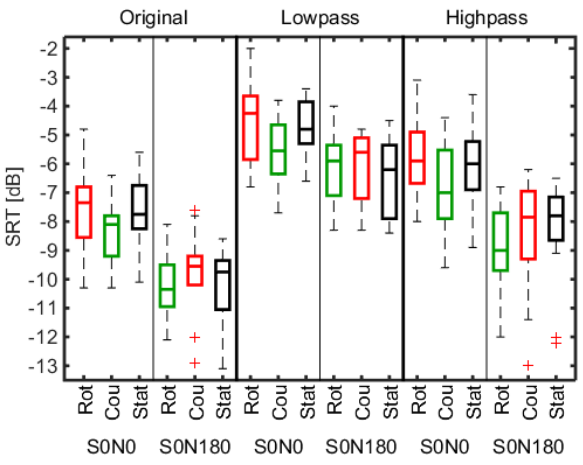

Figure 2: Boxplot of measured SRTs for three filter conditions (original, lowpass, highpass) and three movement conditions (Rotational, Counter-rotational, Static) in the spatial configurations S0N0 and S0N180; green boxes indicate that the sound sources S and N moved towards different ears (diverging ITDs), red boxes mean the sources moved towards the same ear (constantly equal ITDs).

Median RFM values are 0.8 dB (S0N0 and S0N180) in the original condition, 0.9 dB (S0N0) and 0.2 dB (S0N180) in the lowpass condition, and 1.3 dB (S0N0) and 0.4 dB (S0N80) in the highpass condition. A t-test was applied to test whether values differ significantly from zero. For S0N0 all p-values were <0.01 (original condition) or even <0.001 (lowpass and highpass condition), while in the S0N180 condition only the original condition shows a significant RFM with a p-value of p<0.001.

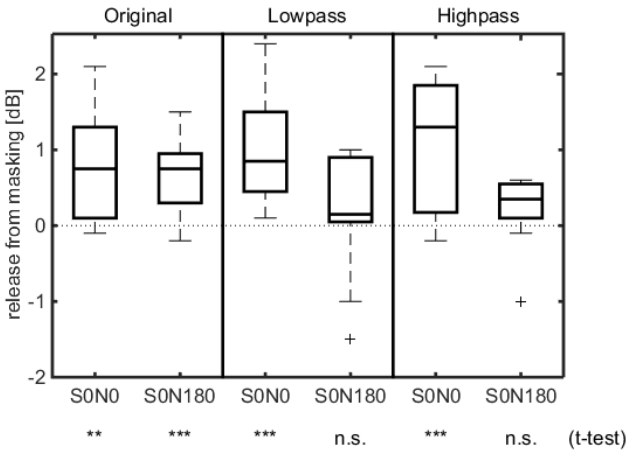

Figure 3: Release from masking by dynamic binaural cues for three filter conditions (original, lowpass, highpass) and the spatial configurations S0N0 and S0N180; the boxplot shows the difference red SRT – green SRT calculated for each of the 12 participants; statistical significance was tested using a t-test (** p<0.01; *** p<0.001).

Discussion

Low- and highpass filtering generally led to higher SRTs (decreased speech intelligibility) than the original condition. Static SRTs reproduced relations between the filter conditions and the spatial configurations S0N0 and S0N180 known from literature [3,5].

The sound source movements led to equal tendencies in the two filtered conditions. Highly significant RFM values were measured in condition S0N0 with lowpass and highpass filtering, whereas in condition S0N180 no RFM was observed in both cases. Only the original condition shows the same amount of RFM for S0N0 and S0N180. Thus, the amount of RFM is not related to the quantity of front-back confusions, which is higher for lowpass filtered signals [4]. This indicates that dynamic unmasking of speech and resolving front-back confusions are independent effects. Furthermore, it can be stated that interaural level differences are as important for the dynamic unmasking as the ITDs.

The lack of RFM in condition S0N180 low and high may be explained by the pinna effect, which may cause a sufficient SNR improvement that the minimal angular changes are not needed for intelligibility. Although ITDs are not affected, pinna cues might be relevant at frequencies close to 1 kHz in the lowpass condition. These are important for consonant recognition when higher frequencies are missing.

Conclusion

The effect of minimal rotational and counter-rotational sound source movements on speech intelligibility was investigated for lowpass and highpass filtered signals and for the unfiltered speech and noise. By testing the conditions S0N0 and S0N180 it was assessed whether the amount of spatial release from masking achievable by movements is related to the quantity of front-back confusions occurring in these configurations. For S0N0 all filter conditions led to similar significant amounts of RFM. This indicates that dynamic unmasking of speech is not related to resolving front-back confusions.

Literature

[1] Begault, D., Wenzel, E., Lee, A. and Anderson, M.: Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. J. Audio Eng. Soc. 49.10 (2001), 904-916

[2] Behrens, T.: Der ‚Kommunikations-Akustik-Simulator' im Oldenburger ‚Haus des Hörens'. Fortschritte der Akustik - DAGA 2005, München, 443-445

[3] Bronkhorst, A.: The Cocktail Party Phenomenon: A Review of Research on Speech Intelligibility in Multi-Talker Conditions. Acustica united with Acta Acustica 86 (2000), 117-128

[4] Brungart, D. and Simpson, B.: Effects of bandwidth on auditory localization with a noise masker. J. Acous. Soc. Am. 126 (2009), 3199-3208

[5] Gilkey, R. and Good, M.: Effects of frequency on free-field masking. Human Factors (37) (1995), 835-843

[6] Grimm, G., Coleman, G. and Hohmann, V.: Realistic spatially complex acoustic scenes for space-aware hearing aids and computational acoustic scene analysis. 16. Jahrestagung der Deutschen Gesellschaft für Audiologie, Rostock (2013), CD-Rom, 4 pages

[7] Grimm, G. and Hohmann, V.: Dynamic spatial acoustic scenarios in multichannel loudspeaker systems for hearing aid evaluations. 17. Jahrestagung der Deutschen Gesellschaft für Audiologie (2014a), CD-Rom, 4 pages

[8] Grimm, G., Wendt, T., Hohmann, V. und Ewert, S.: Implementation and perceptual evaluation of a simulation method for coupled rooms in higher order ambisonics. Proceedings of the EAA Joint Symposium on Auralization and Ambisonics (2014b), 27-32

[9] Heeren, J., Grimm, G., Hohmann, V.: Evaluation of an ambisonics system for psychoacoustical measurements in none-anechoic conditions. Proceedings BMT 2014, 48. Jahrestagung der DGBMT, 3-Länder-Tagung D-A-Ch, Hannover, 859-862

[10] Heeren, J., Grimm, G., Hohmann, V: The influence of dynamic binaural cues on speech intelligibility in headphone and free field listening". Fortschritte der Akustik – DAGA 2015, Nürnberg, 117-120.

[11] Neukom, M.: Ambisonics Panning. Audio Engineering Society Convention 123 (2007), 7297 ff.

[12] Saberi, K. et al.: Free-field release from masking. J. Acoust. Soc. Am. 90 (1991), 1355-1370

[13] Wagener, K., Brand, T. and Kollmeier, B.: Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests. Z Audiol 38 (1999), 4-15

[14] Wightman, F. and Kistler, D.: Resolution of front-back ambiguity in spatial hearing by listener and source movement. J. Acoust. Soc. Am. 105 (1999), 2841-2853