The influence of dynamic binaural cues on speech intelligibility in headphone and free-field listening

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 intelligibility level difference (ILD) is defined as the difference in speech reception thresholds (SRT) between spatially separated speech at 0 deg. and noise at X deg. azimuth (S0NX) and for collocated frontal speech and noise (S0N0). ILDs are largest when the noise is presented from the side and minimal for S0N180. It is known that front-back confusions occurring frequently in localization experiments can be resolved by small head movements that introduce dynamic binaural cues. This study investigates whether these cues may also lead to speech unmasking, i.e., to a higher ILD in the otherwise diotic S0N180 condition. SRT measurements with static and dynamic binaural cues have been conducted for S0N0 and S0N180 in normal-hearing listeners. The stimuli were rendered using 11th order ambisonics. Due to differences between SRTs measured with headphones and SRTs measured with loudspeakers, which may also be attributed to dynamic binaural cues, both reproduction methods were used. Results show that dynamic binaural cues improve speech intelligibility by the same amount for both reproduction methods. A comparison with a Binaural Speech Intelligibility Model (BSIM) shows that the measured SRTs can be reproduced by the model.

Funded by DFG FOR1732

Introduction

The cocktail party phenomenon describes the human ability to understand speech in noisy environments depending on the spatial distribution of the sound sources. One of its basic relations is the intelligibility level difference (ILD). It describes the case of a single speech signal presented from the front (Speech at 0° azimuth, S0) and a single noise sound source for various azimuths (NX). For each of these conditions (S0NX) the speech reception threshold (SRT) can be compared to the SRT of a reference condition (S0N0). Thus, the ILD defines a measure for the spatial release from masking (RFM) of speech-in-noise:

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

According to [4] ILDs are largest when the noise is presented from the side (up to 13 dB for S0N120) and minimal for S0N180 (0-3 dB). This shows that spatial separation of the two sound sources generally leads to an improvement of speech intelligibility except for the front/back case S0N180. Special about this condition is its lack of binaural cues. Due to the equal distances of the sound sources to each ear, there are no interaural time delays (ITD) or interaural intensity differences (IID) that could provide cues to segregate the sound sources. The only cue available is the spectral shape caused by the pinnae, which is a monaural effect. For other noise azimuths (30-150°) the impact of binaural cues to the ILD (binaural intelligibility level difference, BILD) is about 4-5 dB [5].

From localization experiments it is known that front-back confusions may occur. When a masking noise is present front-back confusion rates depend on the SNR [7]. In the S0N180 condition of the ILD measurement signals are commonly presented at SNR of -11 to -7 dB (testing normal hearing listeners). For this SNR range Good and Gilkey [7] observed front-back confusion rates of about 50%, which means complete confusion. These results were measured for a pulsed noise target at 0° or 180° masked by continuous noise at 0°.

Front-back confusions can be resolved by head movements, which introduce dynamic binaural cues [1,16]. This leads to the hypothesis that dynamic binaural cues may also help segregating front and rear sound sources in ILD measurements and lead to an improvement of speech intelligibility in the S0N180 condition. Two versions are conceivable: a continuous stream segregation or a temporal RFM due to lateral displacement. Version one is considered with an enabled BILD effect, which is active during the complete test situation. In the second version only short temporal benefits would occur and it would lead to lower RFM values. A comparison of measured results with model predictions by the Binaural Speech Intelligibility Model (BSIM) [3] could provide some evidence. BSIM offers the opportunity to consider binaural advantages per temporal block. If the model can generally predict the effect of movements correctly, the temporal resolution can be derived.

Investigating effects of movements requires the consideration that these can also deteriorate speech intelligibility [11] and cause localization blur [6]. A distinction between the effect of binaural unmasking and the pure movement effect is necessary. Another problem is the choice of the appropriate presentation method. On the one hand only headphone presentation can display the S0N180 condition in a diotic way and show the effect of adding binaural cues purely, on the other hand the performance of subjects in spatial hearing experiments is significantly more precise when loudspeakers are used [17]. Deviations of the measured values, which occur due to the chosen presentation method, may even be related to the presence or absence of dynamic binaural cues. Using both methods may help clarifying.

Therefore, following hypotheses were investigated:

1. Dynamic binaural cues lead to an improvement of speech intelligibility.

2. The RFM effect by dynamic binaural cues (hypothesis 1) is more significant when headphone presentation is used than for loudspeaker presentation. Movements lead to an approximation of the headphone results towards the results with loudspeaker listening.

3. Results for dynamic RFM can be used to improve the Binaural Intelligibility Model by deriving appropriate parameter settings.

Method

Approach

The condition of interest S0N180 as well as the reference condition S0N0 were tested for three movement conditions:

A) S and N moving corresponding to head movements

B) S moving corresponding to head movements an N in the opposite way

C) without movements

This way movements with constantly equal ITDs for the two sound sources and movements with opposed ITDs are created. A sketch of the movement conditions is shown in figure 1. In condition S0N180 movement B leads to equal ITDs and movement A causes opposed ITDs. For S0N0 it is the opposite. With equal ITDs no binaural advantage is expected [14]. It is expected to show the „pure“ movement effect, which is considered to deteriorate speech intelligibility. Opposed ITDs will cause binaural RFM and may lead to a better speech intelligibility than equal ITDs. By analyzing the difference of SRTs for condition A and B the pure effect of dynamic binaural cues can be considered, without being disguised by movement factors. Due to possible differences between results measured with headphones and loudspeakers, which may be related to head movements or dynamic binaural cues, the measurement was conducted for both reproduction methods.

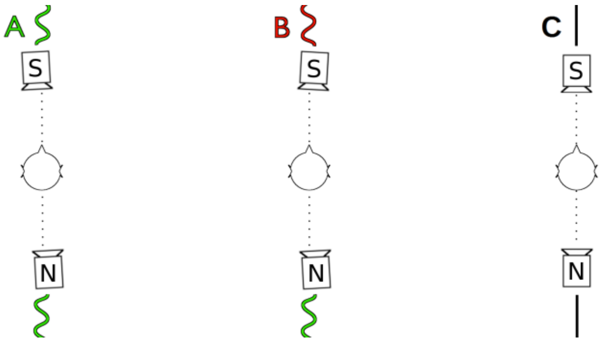

Figure 1: Sketch of the three movement conditions for S0N180; source movements were implemented as a modulation of the nominal azimuths, which were either in-phase for S and N (condition A), anti-phase (B) or sources were not moved (C)

Speech-in-noise test

SRTs were measured using the Oldenburg Sentence Test (OLSA) [15]. The speech level was adjusted adaptively towards the 50%-speech reception threshold. Sentences were masked by a stationary background noise called Olnoise, which is the corresponding speech-shaped noise of the OLSA. It was presented at the fixed level of 65 dB spl.

Virtual source movements were implemented as a modulation of the nominal azimuth by a sinusoid (1 Hz, 10°). The modulation phase was either equal for the speech and noise azimuth (condition A, corresponding to head movements) or opposed (condition B). In condition C the sources were not moved. SRT measurements for the three movement conditions (A, B, C) were tested in interleaved order. Thus, a measurement run consisted of 3x20 sentences and was performed for two spatial configurations (S0N0, S0N180).

Setup

Speech perception tests were performed using two presentation methods: a horizontal loudspeaker array and headphones (Sennheiser HDA200). The loudspeaker array consisted of 24 Genelec 8020 monitors set up regularly on a circle with a radius of 2 m. It was placed in a sound treated room (Communication Acoustics Simulator, House of Hearing, Oldenburg), which has a reverberation time of 0.4-0.6 s (T60) across all frequencies [2]. The direct sound of all loudspeakers was compensated regarding phase delay and spectral shape in the range of 400-22050 Hz. For this purpose impulse responses of all loudspeakers were recorded with a Neumann KM183 microphone that was placed in the center of the loudspeaker setup. The lower cutoff frequency of the compensation is caused by the time difference between direct sound and the first reflection (2.4 ms). An eleventh order basic ambisonics algorithm [13] was used for the panning of the virtual sound sources on the loudspeaker array as well as on headphones. It is part of the Toolbox for Acoustic Scene Creation and Rendering (TASCAR) [8,9,10]. Headphone signals were created by a convolution of the 24 output signals of the panner with head related transfer function (HRTF) of the appropriate loudspeakers. HRTFs were recorded with an artificial head by KEMAR, placed on a chair in the center of the loudspeaker setup. During the measurements participants were placed in the same position. They were told to look at a fixed point at 0° azimuth and keep their heads still. The setup was evaluated by measuring the perceptual spatial resolution, which is 2.7° minimum audible angle for the Olnoise stimulus at 0° on both reproduction methods [12].

Binaural Speech Intelligibility Model (BSIM)

In order to reproduce the measured data with BSIM a Matlab implementation of the model according to [3] was used. As the data is based on spatially dynamic situations BSIM was used in the short time mode. The relations between movement conditions were not modelled appropriately when the recommended EC error processing was active, which is an essential stage for fitting standard ILD data. Thus, the program was run without using the EC error stage. After certain fitting iterations it was found that a blocksize of 28000 samples leads to the best match regarding the relation of movement conditions (at a sample rate of 44100). BSIM values were normalized to the measured median value for S0N0$_C$ with headphone listening.

Participants

Twelve normal hearing subjects participated in the measurements (seven male, five female, age: 21-42 years). All of them had experiences in speech-in-noise tests.

Results

Figure 2 shows a boxplot of SRTs with medians and interquartile ranges. The data in the first column was measured using loudspeakers, the second column shows data measured with headphones. Additionally, the headphone column contains modelled SRTs calculated by BSIM. In the static reference condition S0N0$_C$ the median SRTs are -7.8 dB for loudspeaker presentation (LS) and -6.7 dB for stimuli presented using headphones (HP). Condition S0N0$_B$ led to median SRTs of -8.1 dB (LS) and -8.0 dB (HP). For S0N0$_A$ SRTs of -7.4 dB (LS) and -7.5 dB (HP) were measured.

The median values for S0N180$_C$ are -9.8 dB (LS) and -8.4 dB (HP). With moving sound sources SRTs of -10.4 dB (S0N180$_A$-LS), -8.7 dB (S0N180$_A$-HP), -9.6 dB (S0N180$_B$-LS) and -8.1 dB (S0N180$_B$-HP) were measured.

The BSIM results are median SRTs for five selected OLSA sentences. These were selected, because they showed a maximum deviation of 0.3 dB from the headphone medians in condition C at an early state of parameter fitting. Medians for BSIM condition C are -6.7 dB at S0N0 and -8.0 dB at S0N180. For condition A the median predictions are -7.2 dB (S0N0) and -8.5 dB (S0N180), whereas condition B resulted in values of 8.1 dB (S0N0) and -7.8 dB (S0N180).

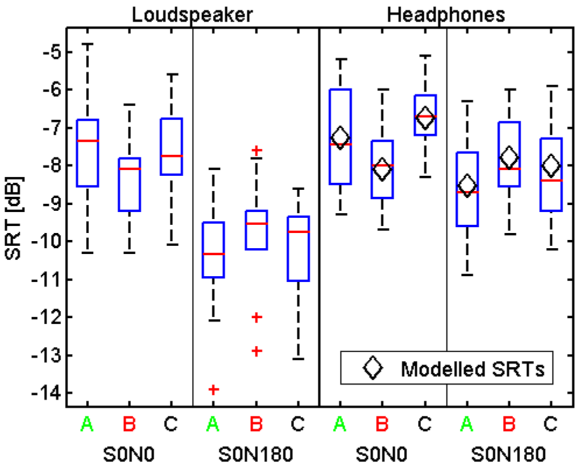

Figure 2: Boxplot of measured SRTs for 12 participants in headphone and loudspeaker listening and modeled SRTs (BSIM) against movement conditions (A,B,C) for S0N0 and S0N180

In figure 3 the release from masking by dynamic binaural cues is displayed. The difference of SRTs for movement condition A and B was calculated for each participant. Medians and interquartile ranges of the absolute values are plotted. The plot is divided in two columns for loudspeaker and headphone presentation. In the conditions S0N0-LS, S0N180-LS and S0N180-HP the median value is 0.8 dB. For S0N0-HP it is 0.7 dB. Significance of the RFM values was checked by a t-test. All measured values are statistically significant (S0N0: p=0.003; S0N180: p<0.001 for both sound reproduction methods).

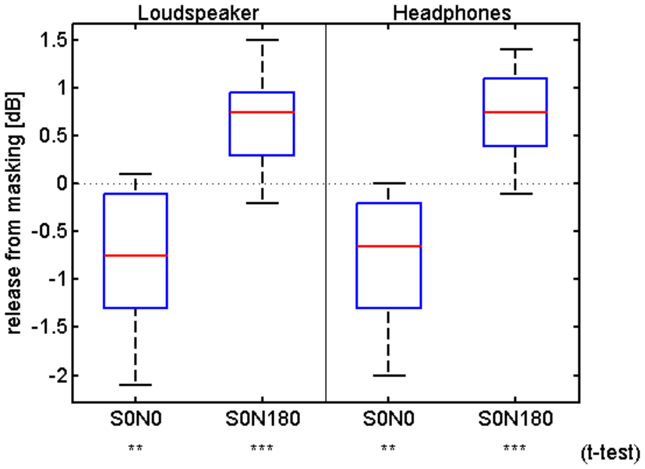

Figure 3: Release from masking by dynamic binaural cues in headphone and loudspeaker listening for S0N0 and S0N180; the boxplot shows the differences SRT$_B$ – SRT$_A$ calculated for each of the 12 participants; statistical significance was tested using a t-test (** p<0.01; *** p<0.001)

Discussion

The results for the static condition C are in line with the literature data according to [4]. As expected, participants performed best in the free field condition, whereas the headphone condition led to slightly higher SRTs. The difference amounts 1.1 dB at S0N0 and 1.4 dB for S0N180. Movements also showed the expected effects. A significant release from masking of 0.8 dB for movements with dynamic binaural cues (S0N0B, S0N180A) compared to movements with constantly equal ITDs (S0N0A, S0N180B) was observed. Thus, the hypothesis that dynamic binaural cues lead to an improvement of speech intelligibility is confirmed. The effect occurred by the same amount for both presentation methods. This does not support the hypothesis that deviations between loudspeaker and headphone listening are caused by head movements (hypothesis two). Otherwise movements would have led to an approximation of the results of the two presentation methods.

To figure out whether the RFM by dynamic binaural cues is based on a continuous stream segregation or on temporal binaural benefits, two aspects were considered. First it has to noted that the RFM is much lower than BILD values according to [5]. Second, BSIM is capable of predicting the small differences between SRTs with a high precision. Both findings do not support the idea of a continuous stream segregation. Contrarily, the results validate the EC theory, which is using temporal binaural benefits. For the fitting of the BSIM results to the measured data it was necessary to apply unusual parameter settings. The large blocksize used for the short time processing provides evidence for a sluggishness in the usability of dynamic binaural cues for speech perception. Additionally, it was necessary to switch off the EC error processing, which is jittering the ITDs and leads to a better reproduction of measured SRTs for noise azimuths of 15-165° [3]. Results imply that there is a need to adjust this feature for an improved modelling of the azimuth dependent perception of interaural cues. This confirms the hypothesis that the results can be used to improve the Binaural Intelligibility Model.

Conclusions

The influence of dynamic binaural cues on speech intelligibility can be explored by the introduced method. Dynamic binaural cues lead to an improvement of speech intelligibility. In headphone and free-field listening the same effect was observed. Results do not support the hypothesis that deviations between headphone and free-field listening can be explained by head movements. The effect can be modelled by the Binaural Speech Intelligibility Model. Further adjustment of the model is necessary.

Literature

[1] Begault, D., Wenzel, E., Lee, A. und 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] Beutelmann, R.: Modelling binaural speech intelligibility in spatial noise and reverberation for normal-hearing and hearing impaired listeners. Dissertation, Universität Oldenburg (2008)

[4] 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

[5] Bronkhorst, A. und Plomp, R.: The effect of head induced interaural time and level differences on speech intelligibility in noise. J. Acoust. Soc. Am. 83 (1988), 1508-1516

[6] Chandler, D. und Grantham, D.: Minimum audible movement angle in the horizontal plane as a function of stimulus frequency and bandwidth, source azimuth, and velocity. J. Acoust. Soc. Am. 91 (1992), 1624-1636

[7] Good, M. und Gilkey, R.: Sound Localization in Noise: The effect of signal-to-noise ratio. J. Acoust. Soc. Am. 99 (2) (1996), 1624-1636

[8] Grimm, G., Coleman, G. und 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

[9] Grimm, G. und 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

[10] 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

[11] Hansen, M. et al.: Speech intelligibility in realistic listening situations for different numbers, azimuths and movement of speech or noise maskers. Proceedings of the International Conference on Acoustics AIA-DAGA 2013, 425-427

[12] 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

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

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

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

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

[17] Yost, W., Dye, R. und Sheft, S.: A simulated "cocktail party" with up to three sound sources. Percet. Psychophys. 58 (1996), 1026-1036