

Evaluation of Local Acoustical Privacy Protection in Small, Enclosed Compartments

Christian Blöcher¹, Alois Sontacchi¹, Thomas Hatheier²

¹ Institute of Electronic Music and Acoustics, Graz, Austria, Email: bloecher@iem.at

² Audio Mobil Elektronik GmbH, Braunau – Ranshofen, Austria, Email: thomas.hatheier@audio-mobil.com

Introduction

Acoustic communication in small, enclosed compartments presents a challenge in establishing private spheres for single users without heavily affecting the acoustical environment for others. In order to reduce speech intelligibility, active acoustical privacy applications typically feature the playback of masking sounds via one or multiple loudspeakers focused on the would-be listener's position. With a microphone in close proximity to the speaker, the masking sounds can be spectro-temporally adapted to the target speech signal in real-time for improved performance [1].

A recently developed acoustical privacy protection system for use in car compartments focuses on applying masking measures *locally* to prevent additional loudness in the car interior. Two explicit and relevant scenarios are displayed in figure 1. All car seats are equipped with active headrests with a built-in pair of loudspeakers and microphones. The speech signal is captured by the microphones in the speaker headrest and used as input for a real-time capable masker generation algorithm. The generated masking signals are then played back via the loudspeakers in the listener headrest. The approach utilizes a dual masking strategy: Firstly, a *broadband noise masker* is spectro-temporally adapted to the target speech and spatialized to the direction of the target speaker relative to the listener to prevent spatial unmasking [2]. Secondly, temporally compact sounds that are spectrally adapted to the target speech are randomly spatialized and played back during specific parts of the target speech. These *distractors* aim to divert the listener and hinder them from comprehending the speech content. To quantify potential benefits of using distractors in loudness increase and user acceptance a listening experiment was conducted. Simulating a real-world scenario, a situation was staged, in which subjects overheard one end of a phone call while obstructed by the privacy application with the task of echoing understood speech contents. Masking levels were adjusted adaptively based on subject responses evaluated by an automatic speech recognition (ASR) system to find individual minimum masking levels for various experiment conditions while ensuring unintelligibility.

Experiment Setup

To ensure control over environmental variables while providing the listeners a comfortable situation, the experiment was carried out in an anechoic chamber. Two configurations for speaker and listener positioned behind each other were investigated, as displayed in figure 2.

To acoustically simulate the driving environment, a circular arrangement of eight Genelec 8020B loudspeakers

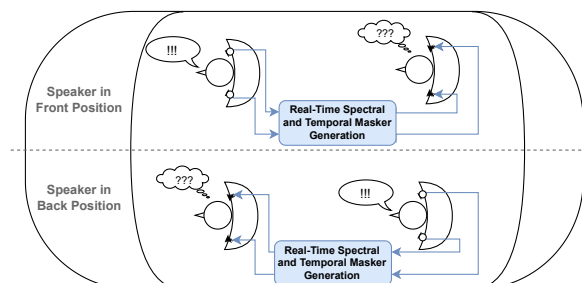


Figure 1: Joint depiction of two application scenarios for in-car local acoustical privacy protection with speaker, listener, and active headrests.

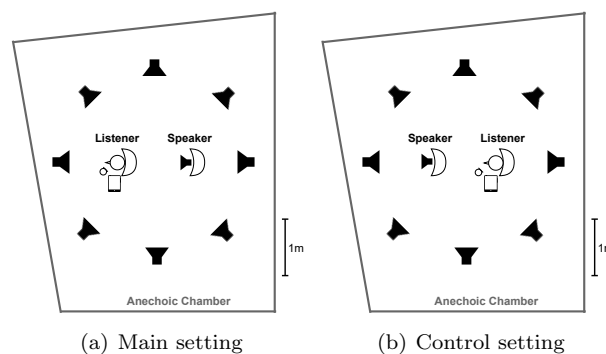


Figure 2: Experiment configurations in anechoic chamber with ambient loudspeaker arrangement, simulated speaker, headrests, listener headset microphone, and tablet device.



Figure 3: Speaker and listener setup.

with a diameter of 3 m was used. Speaker and listener positions were equipped with active headrests and located off-center at a distance of 1 m to each other and the closest ambient loudspeaker on the speaker-listener axis. The speaker was simulated using the front-facing channel of a spherical loudspeaker array as shown in figure 3(a). Subjects sat in a comfortable position with their head leaned back against the headrest. Subjects were provided with a tablet device to give inputs and control the experiment's progression and were fitted with a headset microphone to record speech responses (cf. figure 3(b)).

The experiment procedure was implemented in a Python

backend, which handled the experiment progression, selecting and playing speech samples, and recording and processing subject responses. Real-time masker generation was performed by a Matlab instance. The generated masking audio was routed through and leveled within a Reaper instance, which also contained the pre-compiled eight-channel ambient tracks looped seamlessly for playback. The tablet’s user interface was created using MobMuPlat [3]. Programs communicated with each other through Transmission Control Protocol (TCP) connections and OpenSoundControl (OSC) [4].

Ambient driving noises were simulated for 5 km/h, 30km/h, and 100 km/h, providing an increasing degree of natural masking with higher velocities. Mono in-car recordings were spatialized as described in [5] using the IEM Plugin-Suite’s¹ Granular Encoder [6] and ALLRAD Decoder [7] for playback via the circular eight-channel loudspeaker arrangement to create a diffuse ambient soundscape at both listening positions. The multi-channel ambient noise recordings were leveled such that the resulting calibrated levels of binaural recordings at both positions using a Brüel & Kjær 4128C head and torso simulator (HATS) matched those of calibrated reference in-car HATS recordings (cf. table 1).

Speech samples for the listening experiment were extracted from the Austrian-German Graz Corpus of Read and Spontaneous Speech (GRASS) conversational speech section [8], which provides natural conversation patterns in high quality audio recordings separable by speaker. 34 sets of ten short sentences spoken by male and female speakers were extracted from one dialogue half each for use in the experiment to simulate the situation of hearing only one end of a phone call. Sentences were chosen to not contain names or inappropriate content. To ensure that all sets have a discernable overarching thematic context that can be inferred from the speech contents while still being short enough to be echoed easily by subjects, some extracted samples were edited. The resulting signals had a maximum length of 4 s. All audio files were normalized and the respective sentences transcribed manually. The transcriptions were preprocessed in Python to enable later comparison of *semantic content* with subject responses. Lemmatization was performed using the pretrained spaCy [9] pipeline `de_core_news_sm` and added on postfiltering using a lookup table from [10] with lookups for few misidentified lemmas and lexical equivalents of numerals added manually. Stop words were filtered using a list provided by NLTK [11], which was modified with few corpus-specific word additions and deletions. To account for the *Lombard Effect* speech levels in a distance of 1 m from the simulated speaker were adjusted based on ambient noise levels within the specifications of [12] (cf. table 1). Speech levels for the front speaker position were raised slightly to account for the loudspeaker’s directivity and the absence of reflections e.g. caused by the windshield in real vehicle interiors.

Table 1: Ambient and speech levels per speaker position.

ambient condition	ambient level	speech level	
		back	front
5 km/h	47 dB(A)	56 dB(A)	57 dB(A)
30 km/h	58 dB(A)	60 dB(A)	61 dB(A)
100 km/h	65 dB(A)	63 dB(A)	64.5 dB(A)

Procedure

Masking levels necessary for subjects to understand only 50% of speech contents and resulting subject annoyance were evaluated for standalone broadband masking (BB) and for BB in combination with each of three distractor sound variants *A*, *B*, and *C*. When adding distractors, BB levels were reduced by 3 dB, 6 dB, and 10 dB, resulting in ten different masker conditions. These ten masker conditions were evaluated under each of the three ambient conditions. Ambient, distractor, BB reduction, and set order were shuffled per subject. Due to the experiment duration per configuration of 1 h to 1:30 h, participants (22 normal-hearing native German speakers) were split into two groups by position. Because the back listening position provides the more difficult configuration for speech understanding due to the directivity of the loudspeaker and lack of reflective surfaces within the anechoic chamber, this configuration was used as a control setting. 18 subjects performed the experiment in the main setting, while four subjects participated in the control setting.

To test the functionality of the Google Cloud Text-to-Speech [13] ASR system as a vital component of the experiment and to give subjects a chance to practice, subjects had to echo three sentences of a set without masking in ambient condition 5 km/h while instructed to speak slowly and clearly. The set of sentences used was discarded for the subsequent experiment.

Macro Structure The following was repeated for all ambient conditions. First, subjects were presented with looped sentences of a set masked by the standalone BB and adjusted the BB level to the point where they failed to understand the speech content. Since exposure to looped speech content tends to lead to overestimation of necessary masking levels, and because subjects were unfamiliar with the thematic context of the following set, subject-set levels were reduced by 3 dB. With this starting level an *individual adaptation procedure* was started to determine the true standalone BB masking level at which subjects understand only 50% of speech contents. Subjects then rated the masking signal’s annoyance on an 11-point Likert scale [14]. Subsequently, BB levels were reduced by 3 dB, 6 dB, and 10 dB and levels for each distractor variant *A*, *B*, and *C* were automatically adjusted, so that, based on the previously played sentence, combined masking by BB and distractor resulted in the same loudness increase over ambient noise and speech as the standalone BB component without level reduction. Computation of loudness at the listener position was based on prerecorded ambient noise and spatialized speech and masking signals set to current masking levels. Loud-

¹Available: <https://plugins.iem.at/>

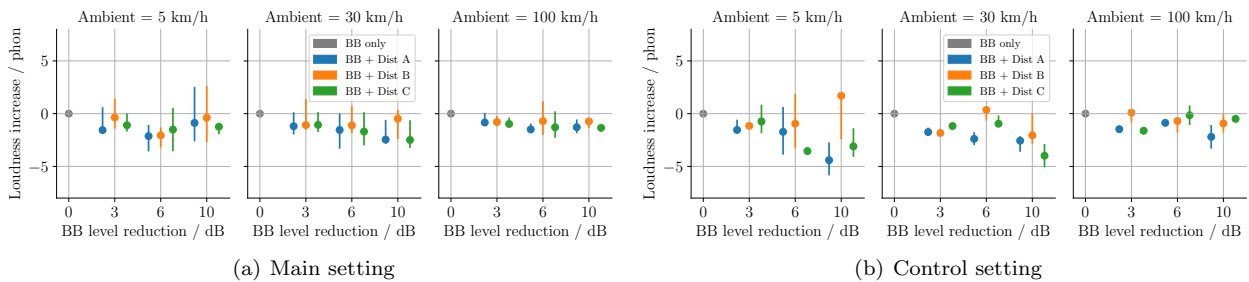


Figure 4: Median values and IQR of mean intra-subject loudness increase at turning points of distractor masking conditions (BB @ $\{-3, -6, -10\}$ dB + Dist $\{A, B, C\}$) over standalone broadband masking (BB @ 0 dB).

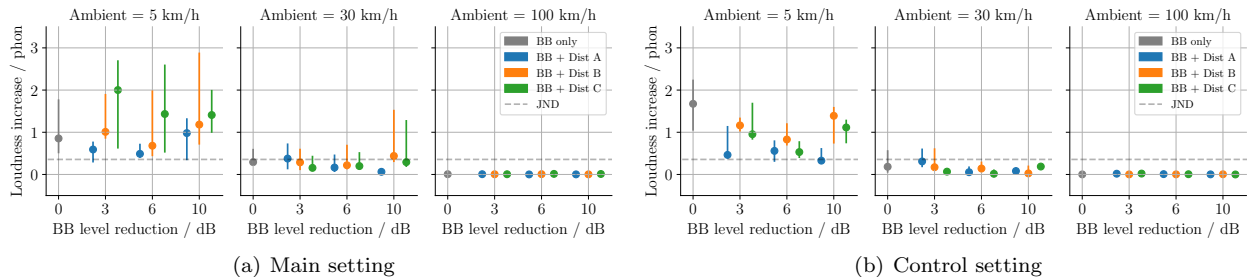


Figure 5: Median values and IQR of mean loudness increase at turning points of standalone broadband masking (BB @ 0 dB) and distractor masking conditions (BB @ $\{-3, -6, -10\}$ dB + Dist $\{A, B, C\}$) over ambient loudness versus JND.

ness computation was performed using the MoSQITo implementation of Zwicker’s loudness model for stationary sounds [15]. Similar to the standalone BB component, with this initial level, individual distractor levels achieving speech masking were adaptively determined for each combination of distractor variant and level reduction. In order not to break subjects’ immersion, ambient noise was present throughout the procedure. After evaluation of all masking conditions for the current ambient condition, subjects could leave the anechoic chamber during a mandatory break of five minutes.

Individual Adaptation To determine masking levels at which subjects understand only 50% of the speech content, a one-up-one-down procedure is used, starting with an initial step size of 3 dB. Beginning with the BB or distractor start levels calculated in the previous step, subjects are presented with a set’s speech material one sentence at a time and instructed to echo any understood speech content or use the skip phrase *Keine Ahnung / no idea*. Responses are automatically transcribed by the ASR system, applying detection boosts to reference transcription and skip phrase to prevent mistakenly transcribing homophonic utterances. If the skip phrase is detected, masking levels are too high for the subject to understand any speech content and the same sentence is presented again with masking levels reduced by the current step size. This is repeated for up to three times before presenting the next sentence in the set. If the skip phrase is not used, the transcription is lemmatized and filtered for stop words. Checking preprocessed reference and response transcriptions for synonyms using OdeNet [16], the lemma recall is calculated. If $> 50\%$ of lemmas are recalled the sentence is assumed understood and masking levels are increased by the current step size. Otherwise, masking levels are decreased. After updating masking levels the set’s next sentence is played.

At turning points the step size decreases by 1 dB. The procedure converges at the current masking level if the step size is reduced to 0 dB and the last sentence was not understood. Otherwise, it is reset to 1 dB. The procedure aborts if there are no sentences left in the current set. Additionally, the procedure is aborted if the sentence is not understood but the masking level is ≥ 15 dB below level of speech, indicating that the component’s masking effect is negligible. In case of distractor adaptation the procedure is also aborted if the sentence is understood but the loudness increase over ambient and speech exceeds that of the standalone BB component by ≥ 5 phon, indicating that the distractor component is ineffective for the subject.

Results

Collected data points were unfortunately sparse, with ca. one third of all adaptation procedures not converging. Stronger than anticipated level fluctuation of masking components based on speech content hindered convergence with only ten available sentences per set. To remedy this, we incorporated adaptation results into the result set that had at least two turning points and did not converge due running out of sentences, retrieving approximately 80% of procedures aborted for lack of sentences.

Loudness The resulting total signal loudness was defined as the N_5 value obtained from computing the loudness for time-varying signals according to [15]. To achieve more consistent loudness values during evaluation, instead of comparing values at convergence points of the respective adaptive procedures, we calculated the intra-subject mean of intra-trial loudness differences between distractor masking conditions and BB, as well as between masking conditions and ambient noise at procedure turning points. Although differences are rarely statistically significant (with $\alpha = 0.05$), some general trends can be

observed. At the listener position (cf. figure 4) for the main setting, all median values indicate lower loudness for distractor conditions. Comparing the distractor variants within both settings, at adverse listening conditions sounds *A* and *C* reduce loudness more than sound *B*, if some base masking provided by BB or ambient sound is present, or the speaker is facing away from the listener. This is evident in the main setting at 30 km/h and the control setting at 5 and 30 km/h. The effect is less pronounced at 100 km/h, because natural masking provided by ambient noise likely already was sufficient for some subjects. Resulting loudness increase at the speaker position over ambient noise (cf. figure 5) is low across all settings and conditions. For 30 and 100 km/h median loudness increase is lower than or roughly equal to the *just-noticeable loudness difference (JND)* [17]. For 5 km/h only distractor variant *A* roughly meets the JND with 3 and 6 dB BB level reduction across both settings.

Annoyance Intra-subject differences in annoyance rating of BB and distractor conditions obtained during the experiment did not show any discernable trends. This is most likely because the timespan between annoyance ratings during the experiment may have been too long, preventing subjects from keeping a consistent frame of reference. More consistent annoyance ratings were obtained in an auxiliary experiment, in which two masked sentences spoken by a male and female speaker were rated side-by-side by five participants of the original experiment in the main setting. Masking stimuli were leveled according to mean settings obtained in the original experiment. Similarly to the obtained loudness results, some trends can be observed in the intra-subject annoyance differences between distractor and standalone BB conditions at 5 and 30 km/h. The results indicate that the additional use of distractors *A* and *C* can reduce subject annoyance caused by speech masking if BB levels are reduced moderately by only 3 dB or 6 dB. However, distractor *B* resulted in increased subject annoyance compared to BB for all conditions.

Conclusion

An approach for local acoustical privacy protection in car compartments was evaluated in a listening experiment simulating a real-world application scenario. Two positions for speaker and listener were examined under multiple driving conditions. Using an automated adaptive procedure to find subject-specific speech masking levels, the effect of the use of distractors in addition to the broadband component on resulting loudness increase at listener and speaker positions, and user annoyance was investigated. Due to methodological shortcomings results were rarely statistically significant but trends in the data suggest that using distractors *A* and *C* yields benefits in loudness at the listener position when broadband masking levels are moderately reduced. Loudness increase at the speaker position is perceptually negligible for driving speeds from 30 km/h across all masking conditions and at 5 km/h with moderate reduction in broadband masking when using distractor *A*. This emphasizes the benefits of focusing on local application of masking mea-

asures, e.g. via active headrests. While annoyance ratings were inconsistent with ratings spaced out over the experiment, subject preferences for distractors *A* and *C* could be obtained when rating masking conditions side-by-side. However, to verify the observed trends with statistically significant results, future research is necessary.

References

- [1] Y. Hioka, J. James, and C. I. Watson, "Masker design for real-time informational masking with mitigated annoyance," *Applied Acoustics*, vol. 159, p. 107073, 2020. DOI: 10.1016/j.apacoust.2019.107073.
- [2] J. F. Culling and M. Lavandier, "Binaural unmasking and spatial release from masking," in *Binaural Hearing: With 93 Illustrations*, R. Y. Litovsky, M. J. Goupell, R. R. Fay, and A. N. Popper, Eds. Cham: Springer International Publishing, 2021, pp. 209–241. DOI: 10.1007/978-3-030-57100-9_8.
- [3] D. I. Iglesia, "The mobility is the message : The development and uses of MobMuPlat," in *5th International Pure Data Convention*, 2016. [Online]. Available: <https://danieliglesia.com/mobmuplat/>.
- [4] M. Wright and A. Freed, *Open Sound Control*. [Online]. Available: <https://opensoundcontrol.stanford.edu/index.html>.
- [5] F. Holzmüller and A. Sontacchi, "Frequency limitation for optimized perception of local active noise control," Hamburg, 2023. [Online]. Available: https://pub.dega-akustik.de/DAGA_2023/data/articles/000531.pdf.
- [6] S. Riedel, M. Frank, and F. Zotter, "The effect of temporal and directional density on listener envelopment," *Journal of the Audio Engineering Society*, vol. 71, no. 7/8, pp. 455–467, 2023. DOI: 10.17743/jaes.2022.0088.
- [7] F. Zotter and M. Frank, "All-around ambisonic panning and decoding," *Journal of the Audio Engineering Society*, vol. 60, no. 10, pp. 807–820, 2012. [Online]. Available: <https://www.aes.org/e-lib/browse.cfm?elib=16554>.
- [8] B. Schuppler, M. Hagmüller, and A. Zahrer, "A corpus of read and conversational austrian german," *Speech Communication*, vol. 94, pp. 62–74, 2017. DOI: 10.1016/j.specom.2017.09.003.
- [9] *SpaCy*, version 3.5.1. [Online]. Available: <https://github.com/explosion/spaCy>.
- [10] M. Liebeck and S. Conrad, "IWNLP: Inverse Wiktionary for Natural Language Processing," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Association for Computational Linguistics, 2015, pp. 414–418. [Online]. Available: <http://www.aclweb.org/anthology/P15-2068>.
- [11] S. Bird, E. Loper, and E. Klein, *Natural Language Processing with Python*. O'Reilly Media Inc., 2009. [Online]. Available: <https://github.com/nltk/nltk>.
- [12] "Ergonomics — Assessment of speech communication (ISO 9921:2003)," International Organization for Standardization, 2003.
- [13] *Google Cloud Speech-to-Text*, version v1 de-DE, model latest_long. [Online]. Available: <https://cloud.google.com/speech-to-text>.
- [14] "Acoustics — Assessment of noise annoyance by means of social and socio-acoustic surveys (ISO 15666:2021)," International Organization for Standardization, 2021.
- [15] G. F. Coop, *MoSQITo*, version 1.0.8, 2021. DOI: 10.5281/zenodo.6675733.
- [16] M. Siegel and F. Bond, "OdeNet: Compiling a German Wordnet from other resources," in *Proceedings of the 11th Global Wordnet Conference*, University of South Africa (UNISA): Global Wordnet Association, 2021, pp. 192–198. [Online]. Available: <https://aclanthology.org/2021.gwc-1.22>.
- [17] J. B. Allen and S. T. Neely, "Modeling the relation between the intensity just-noticeable difference and loudness for pure tones and wideband noise," *The Journal of the Acoustical Society of America*, vol. 102, no. 6, pp. 3628–3646, 1997. DOI: 10.1121/1.420150.