Spatial-temporal integration of speech reflections

Jan Rennies[1], Anna Warzybok[2] , Thomas Brand[2], Birger Kollmeier[1,2]

[1] *Fraunhofer Institute for Digital Media Technology IDMT, Institute Branch Hearing, Speech, and Audio Technology, and Cluster of Excellence Hearing4all, 26129 Oldenburg, E-Mail: jan.rennies@idmt.fraunhofer.de*

[2] *Universität Oldenburg, Medical Physics, and Cluster of Excellence Hearing4all, 26129 Oldenburg,*

Abstract

In reverberant rooms speech is reflected at boundaries and objects and superimposes with the direct sound, thus creating a complex pattern of temporally delayed, spectrally modified and spatially distributed copies of the direct sound. The fundamental assumption of standard measures and models employed to predict speech intelligibility is that reflections arriving briefly after the direct sound can be integrated and are, hence, useful for speech intelligibility, while reflections arriving later than about 50 to 100 ms after the direct sound are detrimental. This assumption was challenged in a series of experiments within this study by systematically varying the energetic, temporal and binaural properties of direct sound, the reflections, and a stationary noise masker. Speech reception thresholds were measured in normal-hearing listeners. In conditions where either energy or binaural information favor the early components of the room impulse response (RIR), the data confirm that adding reflections with delays beyond a critical time window cannot be perfectly integrated. However, in conditions where the later RIR components are favorable in terms of energy or binaural information, the auditory system appears to ignore the early components and exploit the late components instead. This cannot be modeled by any current speech intelligibility prediction models.

Introduction

In real rooms, sounds emitted by a source are reflected at boundaries and objects, creating a complex pattern of temporally shifted, spectrally modified and spatially distributed reflections which superimpose the direct sound arriving at the listener’s ears. With respect to speech intelligibility, all current models and instrumental measures assume that only the reflections arriving within a certain temporal window of about 50-100 ms after the direct sound can be (at least partially) integrated and, hence, are useful for speech recognition. In contrast, reflections arriving later, i.e., outside the temporal window, can no longer be integrated and are considered detrimental. This concept of separating room impulse responses (RIRs) into early (assumed useful) and late (assumed detrimental) components is supported by many experimental studies (e.g., [1,2,3]), is the basis of standard room acoustical measures such as ‘clarity’ or ‘definition’ and has also been successfully integrated into more complex binaural speech intelligibility prediction models [4].

There may be conditions, however, in which it is not necessarily the early part of an RIR, which provides the most relevant speech information. For example, when late reflections or echoes have a much higher energy than the direct sound (which may occur, e.g., when the direct sound is amplified and played back with a significant delay and at a high volume via a reinforcement system). Similarly, one could create conditions in which the late reflections carry a binaural advantage (such as an interaural phase difference, IPD) in a given noise, while the direct sound does not. It is unclear if under such conditions listeners still rely more on the direct sounds and the early reflections or if – at some point – the late RIR components “take over” and become the dominant source for extracting speech information.

This was explicitly investigated in this study by considering artificial RIRs which only consisted of the direct sound and a single reflection. The RIRs were designed such that – in some conditions – the reflection was delayed by 200 ms (i.e., it lay way outside the typical integration window), but carried an energetic or binaural advantage relative to the direct sound to test how subjects extracted speech information in such conditions.

In addition, experimental data were compared to predictions of an established model representing the classical early/late separation approach [4] as well as a modified version of this model, which allowed a more flexible extraction of information from the RIRs.

Methods

Participants

Eight native English listeners participated in this study (three female, five male). All had normal hearing and were familiarized with the task and speech material before starting the experiment.

Stimuli

The American English matrix test [5] was used in this study. It contains grammatically correct but predominantly non-sense sentences uttered by a female talker, which consist of five words (name, verb, number, adjective, object). The sentences are not predictable, as for each of the five parts of the sentence 10 possible words exist to fill that respective slot. A stationary speech-shaped noise was used as masker. The noise had the same long-term spectrum as the target speech material.

The noise level was set to 65 dB SPL, while the speech level was varied adaptively to generate the desired SNRs. Stimuli were presented to the listeners via open headphones (Sennheiser HD280 pro).

Procedure and conditions

For each condition, speech recognition thresholds (SRTs), i.e., the SNRs required to understand 50% of the presented target words, were measured using an adaptive procedure.

The task of the subjects was to select the words they had recognized on a computer screen after every sentence. The SNR was then adjusted adaptively to converge to the SRT. For each condition**,** a list of 20 sentences was used.

Three different conditions are presented in this contribution. The first condition tested the influence of relative reflection amplitude on SRTs. To this end, the reflection delay was fixed at 200 ms, and the multiplied by a scaling factor α. Note that the speech level was always rescaled so that the overall level (including the contributions of both direct sound and reflection) was at the desired level. In other words, the level of the direct sound was the smaller, the larger the level of the reflection. For α=1, both direct sound and reflection had the same level, which was about 3dB lower than the level of the direct sound in the direct-sound-only condition (α=0). In this condition, no binaural processing was tested, i.e., direct sound, reflection and noise all had an IPD of 0.

In the second and third condition, α was fixed at a value of 1 (same level as direct sound), and the reflection delay was varied between 10 and 200 ms. In condition 2, there was again no binaural processing since all components had an IPD of 0. In condition 3, the IPD of the reflection was manipulated by multiplying the reflection component of the RIR by -1 at the left ear only. This resulted in an IPD of 180° (π) for the reflection, while direct sound and noise still had an IPD of 0.

Prediction models

Predictions were made using the binaural speech intelligibility model (BSIM) proposed in [4]. It processes speech and noise signals for the left and right ear in a bank of auditory filters. In each filter, an independent equalization-cancelation (EC) mechanism is employed to exploit potential binaural information to enhance the subband-SNRs, which are then fed into the Speech Intelligibility Index (SII), from which predicted SRTs are derived. This model assumes that only the direct sound contributes fully to speech intelligibility, while the contribution of delayed reflections gradually decreases with increasing delay. This is realized by extracting the useful part of the RIR by multiplication with a decreasing ramp. The detrimental part is extracted by a complementary window, and is added to the external masker at the input of the model. This model was very successful in predicting SRTs in various spatio-temporal conditions measured in [4]. Because this model follows the established convention “early = useful, late = detrimental”, it is referred to as BSIM-EL in the following.

In addition, a modified model version was tested. In this version, the extraction of the useful RIR components was not limited to starting at the direct sound, but could be realized flexibly depending on where the maximum degree of binaural information was found. This was realized by implementing a sliding temporal window (triangle in Figure 1) and allowing the model to place the peak of the triangular integration window anywhere along the RIR. As in BSIM-EL, the useful part was then extracted by multiplying the RIR with the temporal window, and the detrimental part was extracted by multiplication with the complementary window. This modified version is referred to as BSIM-UD in the following.

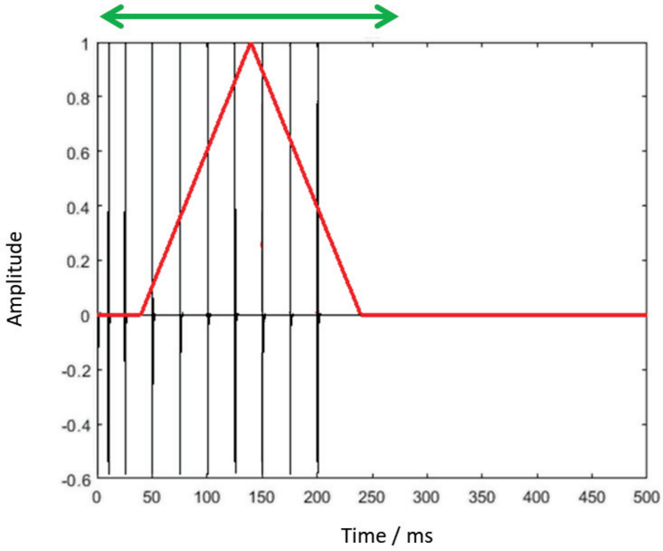

Figure 1: Illustration of the flexible temporal integration window (triangle) employed by the BSIM-UD model to extract the useful part from an RIR.

Results

SRTs measured in condition 1 are shown as a function of reflection amplification in Figure 2. Symbols indicate mean values across subjects, errorbars represent standard errors. As expected, SRTs were lowest in the direct-sound-only condition (α=0). The SRT would have occurred if α had been very high (“inf.”), because in this case again only a single component of the rescaled RIR would have been audible. For low values of α, SRTs were the same. For values of α close to one, SRTs were significantly increased by 3-4 dB. The dashed line in the upper panel shows prediction of BSIM-EL. It is obvious that this model well predicts the SRT pattern for small values of α, but fails tp predict SRTs for higher values of α. This is expected since this model always considers the late reflection (delay 200 ms) to be detrimental, even when its relative energy makes it the dominant source of speech information. The bottom panel of Figure 1 shows predictions of BSIM-UD. The flexible integration of speech information results in a very good match between data and predictions.

Data of conditions 2 and 3 are shown in Figure 3. When there is no binaural information contained in the reflection ($D_0R_0N_0$, shown as circles), SRTs gradually increased with increasing reflection delay, essentially replicating results of [3]. This was well predicted by BSIM-EL. When an IDP was introduced for the reflection only ($D_0R_0N_0$, squares), the SRT pattern changed considerably: SRTs were much lower (by about 5-6 dB) when a reflection was present than in the direct-sound-only condition. This was statistically independent from reflection delay, i.e., SRTs remained low even at delays that are normally detrimental for speech intelligibility. This could not be predicted by BSIM-EL (dashed line in top panel). At the longest reflection delay, the mean SRT was overestimated by about 8 dB.

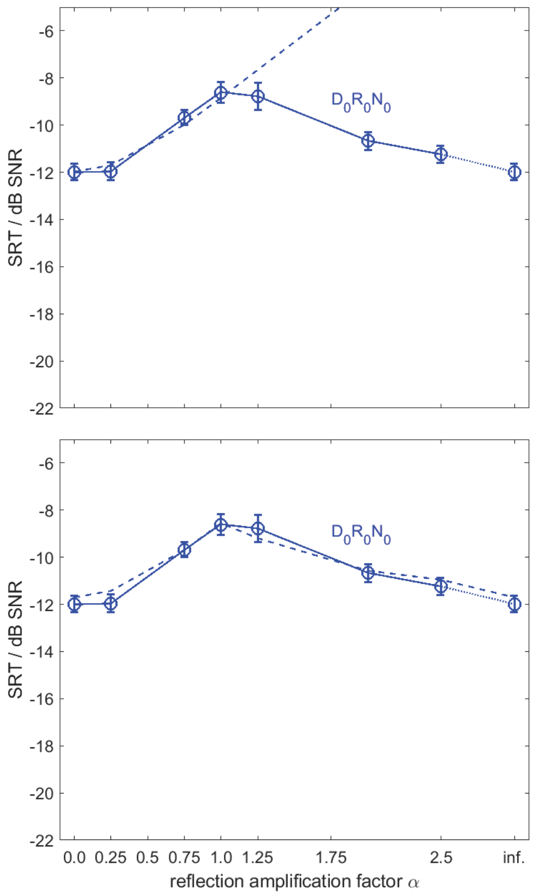

Figure 2: Mean SRTs (symbols) and standard errors measured in condition 1. Dashed lines represent predictions of BSIM-EL (top) and BSIM-UD (bottom).

The corresponding predictions of BSIM-UD are shown as dashed line in the lower panel of Figure 3. Apart from a smaller deviation in condition 2, all SRTs were again quantitatively predicted by the model.

Discussion

The experimental data collected in this study clearly show that the binaural auditory system is capable of focusing on highly delayed components transmitted via an RIR if these components are energetically or binaurally favorable. In other words, it is not always the case that – as commonly assumed – the direct sound and the early reflections determined the useful part of an RIR.

This cannot be modeled by models implementing the classic assumption that late reflections are always detrimental for speech intelligibility. In contrast, the conceptual model approach tested in this study allowed for a full flexibility as to the temporal position of the window to extract the useful part from the RIR. This model could quantitatively predict all conditions measured in this study, further supporting the notion that the auditory system is able to extract temporal information in a highly flexible way. The mechanisms underlying this flexibility are so far unknown. The present BSIM-UD model required full access to the speech signal, the RIR and the noise.

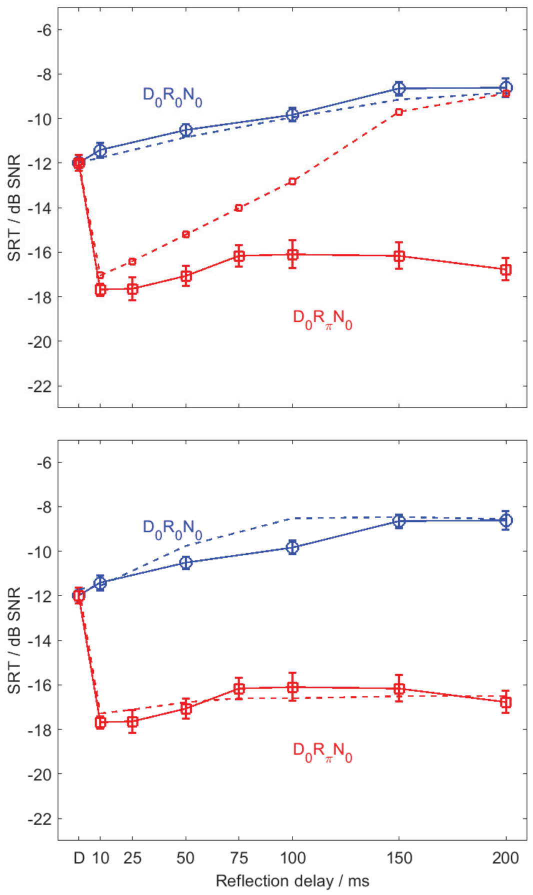

Figure 3: Mean SRTs (symbols) and standard errors measured in condition 2 (circles) and 3 (squares). Dashed lines represent predictions of BSIM-EL (top) and BSIM-UD (bottom).

It is obvious that the auditory system does not have this kind of oracle knowledge, and the mechanisms responsible for the remarkable degree of flexibility in spatio-temporal integration remain subject to future research.

Acknowledgements

This study was supported by the German Research Foundation (DFG) Grant RE 4160/1-1 and Projektnummer 35201583 – SFB 1330 A1, and the National Institutes of Health–National Institute on Deafness and other Communication Disorders Grant No. R01 DC04545. We thank HörTech gGmbH for providing the speech material.

References

[1] Lochner, J.P.A., & Burger, J.F.: The influence of reflections on auditorium acoustics. J. Sound Vibr. 1, 426-454.

[2] Arweiler, A. & Buchholz, J.: The influence of spectral characteristics of early reflections on speech intelligibility. J. Acoust. Soc. Am. 130, 996-1005.

[3] Warzybok, A., Rennies, J., Brand, T. & Kollmeier, B.: Effects of spatial and temporal integration of a single early reflection on speech intelligibility. J. Acoust. Soc. Am. 133, 269-282.

[4] Rennies, J. Warzybok, A., Brand, T. & Kollmeier, B.: Modeling the effects of a single reflection on binaural speech intelligibility. J. Acoust. Soc. Am. 135, 1556-1567.

[5] Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M., Uslar, V., Brand, T. & Wagener, K.: The multilingual matrix test: Principles, applications and comparisons across languages – a review. Int. J. Audiol. 54, 3-16.

[6] Parasuraman, R. & Wilson, G. (2008). Putting the Brain to Work: Neuroergonomics Past, Present, and Future. Human Factors, 50 (3), 468–474.