euronoise 2021
25th to 27th of October
Madeira, Portugal - online

Noise unmasks the masking effect of reverberation on early reflections in the intelligibility of speech

Nicola Prodi[1], Matteo Pellegatti[2], Chiara Visentin[3]

[1,2,3]Department of Engineering, University of Ferrara, Italy
[1]nicola.prodi@unife.it; [2]matteo.pellegatti@unife.it; [3]chiara.visentin@unife.it

Abstract
Changing the ratio between early and late reflections affects the clarity of speech and music and modulates the spatial impression; this latter effect typically happens when the direction of arrival of reflections is varied too. In this mechanism both the overall amount of energy and the correlation at the ears play a role, the latter being crucial for spatial impressions. When noise is added on speech, it is unclear whether and how the spatial characteristics of the source are altered, nor it is clear if this would affect speech intelligibility. In this work impulse responses with specular or diffuse early reflections and two different reverberant tails are used to create virtual sound fields with control of clarity and reverberation. It is shown that in some cases the presence of noise restores spatial percepts of the speech source that are unavailable in the reverberation-only (quiet) conditions. These cues are associated with an improvement in speech intelligibility.

Keywords: scattering, early reflections, spatial percepts, speech intelligibility

1 Introduction

The importance of early reflections has been underscored both for speech intelligibility and for the appraisal of the spatial characteristics of a sound source. In the former case it has been demonstrated since long that early reflections are beneficial for speech intelligibility because the auditory system can integrate them with the direct sound [1]; the conventional limit between useful and detrimental reflections for speech has been set to 50 ms. Based on this statement the acoustical parameters C50 and U50 [2] have come into use to assess the suitability of a given impulse response to ensure speech perception. The way our auditory system integrates early reflections has been investigated also in the recent works [3,4,5]. The temporal integration of early reflections was deemed as a monaural process which was most effective if reflections were co-located with the direct sound [3]. Anyway, some interaction of monaural and binaural processing was also found when detrimental reflections with long delays should be suppressed [4]. Moreover, the phase relationships between direct sound, reflections and noise appear to have a role too. Reflections with a phase congruent with the direct sound are easily integrated but those with a phase congruent with the noise are harder to integrate; when later reflections are dominant the information they convey can be used by the auditory system [5]. As regards the spatial description of sound sources, and their size or *width* in particular, it is mostly associated with the binaural processing which is often modeled as a cross-correlation of the signals at the ears from which binaural quantities are derived [6]. The percept of *distance* seems less influenced by binaural cues, at least in steady conditions [7]. Moreover, in the study of concert halls also the monaural concept of lateral energy fraction was introduced as a way of emphasizing the role of lateral energy; the spatial percept of source *width* could be related to early portions of the lateral sound relative to the overall level, while the so-called *envelopment* was related to the later components. Both were dependent on their

relative level and had mutual influence [8]. From the above it appears that the analysis of impulse responses for predicting speech intelligibility and for describing the spatial characteristics of a sound source are two almost distinct areas of research. In addition, there is a paucity of studies describing the alterations that the auditory image of a speech source undergoes when noise is added, apart few specific studies involving hearing impairment or processing for hearing aids. It is necessary to bridge the gap between the perceptual appraisal of a speech source and its intelligibility, and to understand how the auditory image is built in quiet as well as in noise and how this image is related to speech intelligibility. In this perspective paper [9] investigated the spatial release from masking (SRM) in relation to the image size and found that SRM was reduced and image size increased with the use of hearing-aids. The work [10] tested several sound reproduction setups and focused on the relationship between image size (a more comprehensive definition of *width*) and system's energy spread. This quantity influenced speech intelligibility but not the image size. The dissociation was motivated by the fact that the inter-aural cross-correlation (IACC), taken as a proxy of the image size, did not change between conditions with different energy spread, while speech intelligibility did since it was sensitive to talker-to-masker ratio and binaural interactions. Furthermore, the recent work [11] showed that the auditory image of a speech source in quiet (no noise) was modulated by the type of a single early reflection, either diffuse or specular. In particular with the diffuse reflection the sound source was perceived closer (along with the percept of *distance*) and more focused (according to the percept of *focus*), while its size (related to the percept of *width*) did not change from the specular to the diffuse reflection. This finding added a concept to the previous literature, marking a first distinction between source *width* and source *focus* and remarking that, even when source dimensions are perceived as constant, the source itself can be perceived more or less blurred. When noise was added in the same experiment and speech intelligibility was measured, better scores were obtained for the diffuse reflection; but the spatial percepts were not evaluated in noisy conditions so the association between speech intelligibility and the perceptual qualities of the sound image was only indirect. In a later incremental study the same authors [12] analysed the cases with only three early specular or diffuse reflections and described the trends of the spatial percepts *distance*, *width* and *focus* in quiet; however they did not perform speech intelligibility tests. The present work is conceived as a further step forward in the analysis of spatial percepts in quiet, in noise and of their relationship with speech intelligibility. In particular two reverberation levels are added to a selection of the impulse responses used in [12] with three diffuse or specular early reflections, and the energetic balance between the early reflections and the sound tail is modulated corresponding to two clarity C50 levels. The research questions that the present work addresses are the following:

1) What is the relationship between the spatial percepts and some basic acoustical variables such as clarity, reverberation and with the type of early reflection?
2) What is the distortion, if any, of the above relationship when noise is added and how do the spatial percepts change between quiet and noisy conditions?
3) Is there as association between the spatial percepts and the speech intelligibility in noise?

2 Materials and methods

2.1 Measurement of the direct sound and of the early reflections

The speech stimuli for all the experiments were created by convolving anechoic speech material with head-related impulse responses (HRIRs) which were obtained by mixing a direct sound and three early reflections measured in an anechoic chamber with gaussian uncorrelated sound tails obtained numerically. The measurements of the direct sound and of the single early reflections were conducted in two configurations (specular and diffuse), by varying the surface placed on the floor of the large anechoic room. Plywood panels were used for the specular configuration while one-dimensional quadratic residue diffusers (QRD) were used for the diffuse configuration. Figure 1 shows the layout of source, receiver and surface providing the reflection. The receiver was either positioned with its left ear oriented towards the test surface or upside down. The former setup was designed to simulate the case of a single reflection reaching a listener's left ear from a side wall; the latter set up was designed to simulate a single reflection reaching the listener's ears

from the ceiling. The geometry was varied to obtain different reflection azimuths (α_r). A sound source with the directivity of a person talking (GRAS44AB) was placed close to the edge of the test surface. A B&K type 4100 head-and-torso simulator was suspended over the test surface, aligned with and facing the sound source.

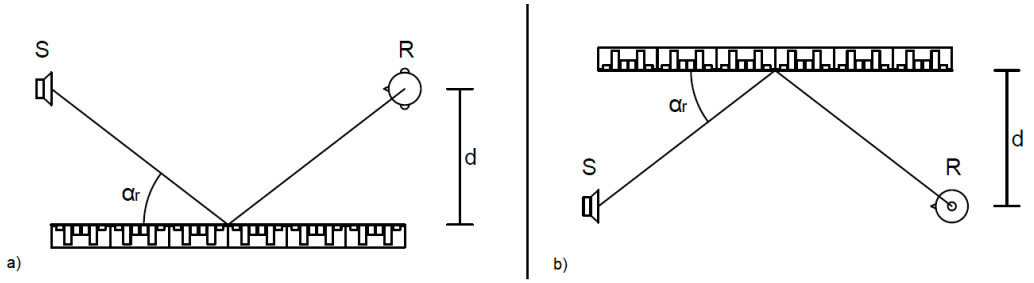

Figure 1 – Schematics of the layouts providing (a) a lateral reflection (upper view), and (b) a ceiling reflection (side view). S-R is the source-receiver distance, αr the angle of reflection, and d the distance from the surface.

The HRIRs with the reflection coming from the right side of the listener were obtained by swapping the left and the right channels of the measured HRIRs. A frontal HRIR (0° azimuth and elevation), including only the direct sound, was measured in fully anechoic conditions and was taken as the reference direct sound for all listening conditions. The measured HRIR were time windowed to retain only the reflection and its total energy was set to -1.6 dB with respect to the direct sound. The early reflections included in the HRIRs used for later convolutions are listed in Table 1. After the direct sound one finds a left reflection at 5 ms from 34° incidence, a ceiling reflection at 8 ms with 34° incidence and a right reflection at 13 ms from 45° incidence. One set was prepared with specular reflections and another set with diffuse reflections.

Table 1 – Direction, angle and timing of the three early reflections included in the HRIRs.

Left wall (Δt = 5 ms)	**Right wall (Δt = 13 ms)**	**Ceiling (Δt = 8 ms)**
✓ (α_{r1} = 34°)	✓ (α_{r3} = 45°)	✓ (α_{r2} = 34°)

2.2 Acoustical conditions

Once the early reflections were available after direct measurement, the reverberant tails were achieved by numerical simulation and mixed in the HRIRs. To do so, first a stereo file was created with incoherent Gaussian noise in the left and right channels. Two types of energetic decays were modelled by simple exponential functions to obtain reverberation times of 0.45 s and 0.85 s (average from 500 Hz to 4kHz). Secondly, the two reverberant tails were mixed with the early reflections with a time delay of 50 ms from the direct sound; the energy ratio between the early sound and the later tail was manipulated.

Table 2 – Measured values of acoustical indicators for the HRIRs used in the auralizations. Data are the averages of left and right value. Rt is reverberation time T20, EDT is the early decay time, C50 is clarity for speech, U50 is the useful-to-detrimental ratio and STI is the speech transmission index, $IACC_{E3}$ is the inter-aural cross-correlation on 80ms averaged over 0.5 - 2kHz. Data for Rt, EDT, C50 are averages 0.5 – 4kHz, U50 is from C50(500Hz-4kHz). Differences between left and right ear were below JND.

	diffuse						specular					
	Rt [s]	EDT [s]	C50 [dB]	U50 [dB]	$IACC_{E3}$	STI	Rt [s]	EDT [s]	C50 [dB]	U50 [dB]	$IACC_{E3}$	STI
Rt1_C1	0.43	0.83	2.7	-8.2	0.61	0.25	0.43	0.86	2.5	-8.3	0.54	0.25
Rt1_C2	0.48	0.87	8.6	-6.7	0.62	0.28	0.49	0.90	8.4	-6.7	0.54	0.28
Rt2_C1	0.82	0.98	2.7	-8.2	0.61	0.24	0.82	1.01	2.6	-8.3	0.54	0.24
Rt2_C2	0.86	0.98	8.7	-6.7	0.62	0.28	0.87	0.99	8.5	-6.7	0.54	0.28

This gave rise to two clarity (C50) values respectively close to 2.6 dB and 8.6 dB (average from 500 Hz to 4kHz). All in all, the acoustical conditions were four (2 Rt x 2 C50) for both specular and diffuse early reflections. Finally, the HRIRs were convolved with anechoic target signals, and the auralized material was presented against a steady state masker with the same spectrum of the target signal (SNR=-6 dB). By doing so two values of U50 were obtained, respectively close to – 8.2 dB and -6.7 dB (average from 500 Hz to 4kHz). Table 2 reports the acoustical parameters for all of the acoustical conditions used in the subsequent listening tests. The data are the average of the left and right ear value but the discrepancies were always less than the respective JNDs. In Expt. 1 listeners were asked to rate a set of spatial percepts (see. Par. 2.3) in quiet (no noise); in Expt. 2 the same spatial percepts were assessed in noisy conditions and in Expt. 3 speech intelligibility scores were obtained in the same acoustical conditions as in Expt. 2.

2.3 Setup of the listening tests

The HRIRs were convolved with the anechoic stimuli of the Word Sequence Test, which consists of sequences of four disyllabic words, preceded by a carrier phrase. For Expt. 1 and 2, five sequences were randomly extracted from the test corpus, for a total duration of 18 s. For Expt. 3, eight test lists were used (2 Rt x 2 C50 x 2 early reflection types); each list was composed of 12 sequences (trials). All stimuli were reproduced using a binaural rendering system installed in a silent room and surrounding the listener. The listener used a touchscreen placed right in front of her/him to input the responses. The speech was presented at a level of 57 dB(A), measured with reference to the participant's left ear. Twenty-one participants took part in Expt. 1, while twenty-five took part in Expt. 2 and 3 (same panel in both experiments). All participants were native Italian speakers and were recruited from among the students and the academic staff at the local university. None of them had extensive experience of listening tests. Prior to the experiment, participants completed a self-administered hearing screening using the IOS device-based application. All participants had test results in the "normal hearing" category (up to 25 dB HL) for the frequencies being tested (0.5-8 kHz). For each trial in the Expt. 1 (quiet) and 2 (noise), participants listened to the playback of the same speech signal. After the audio offset, four visual analog scales (VAS) appeared on the touchscreen and participants were asked to assess the following spatial percepts:
(i) **distance** from the speaker (*How far from you would you locate the speaker?*). The response was given on a VAS with the ends labelled as 1 m and 4 m. For ease of scoring, labels corresponding to 2 and 3 m were also included, together with ticks in 0.2 m steps.
(ii) **focus** of the sound source (*How focused would you rate the sound source?*). The response was given on a VAS with the ends labelled as "very little" and "very much".
(iii) **width** of the sound source (*How wide would you rate the sound source?*). The response was given on a VAS with the ends labelled as "very narrow" and "very wide".
(iv) **envelopment** of the sound source (*How would you rate the sense of immersion in the sound field?*). The response was given on a VAS with the ends labelled as "lower" and "higher".
The percepts are referred to from now on as *distance*, *focus*, *width* and *envelopment*. They were carefully explained to participants before starting the experiment, by means of written instructions. The raw ratings of the spatial percepts underwent a Z-score transformation after pooling all the data, for all his/her tested conditions, by participant. The transformation eliminated between-subject differences while preserving between-condition ones. It hasto be remarked that, thanks to the experimental design and to this normalization procedure, it was also possible to directly compare conditions across blocks. In Expts. 1 and 2 a linear regression model was run for each of the four dependent variables (*distance*, *focus*, *width* and *envelopment*). For the statistical analyses, each model included the following fixed effects: reflection type (diffuse, specular), reverberation (Rt1, Rt2), and in Exp. 1 clarity (C1,C2) while in Expt. 2 useful to detrimental ration (U1, U2). The two- and three-way interactions were included as well. The Expt. 3 was a speech intelligibility measure. One list was used for each condition and the order of lists was counterbalanced across conditions and participants. The scoring was word-based: for each sequence of four words, SI was defined as the proportion of words correctly recognized. In the data analysis of Expt. 3 a generalized linear mixed model was used with the dependent variable speech intelligibility; the fixed effects and interactions were reflection type (diffuse, specular), reverberation (Rt1, Rt2), useful to detrimental ratio (U1, U2) and their two- and three-way interactions. All analyses were conducted with the *R* software, setting

the statistical significance threshold at 0.05. Post-hoc tests and the calculation of the standardized effect sizes (corresponding to Cohen's *d*) were performed with the *emmeans* package. To control for the Type I errors in the case of multiple comparisons, the *p*-values were adjusted using the False Discovery Rate procedure.

3 Results of Experiment 1: spatial percepts in quiet conditions

The results for the assessment of spatial percepts in the quiet conditions are reported in Figure 2. When *distance* was analyzed the main effects of Rt [F(1,198)=26.80, $p<0.001$] and C50 [F(1,198)=39.71, $p<0.001$] were significant. The effect of the type of reflection and all the interactions were non-significant. The post-hoc comparisons revealed the perception of a farther distance of the source for longer Rt (*Rt2 > Rt1*, $p<0.001$, *t.ratio*=5.16, *d*=0.719 – *medium/large*) and for lower C50 (*C1 > C2*, $p<0.001$ *t.ratio*=6.3, *d*=0.878 - *large*).

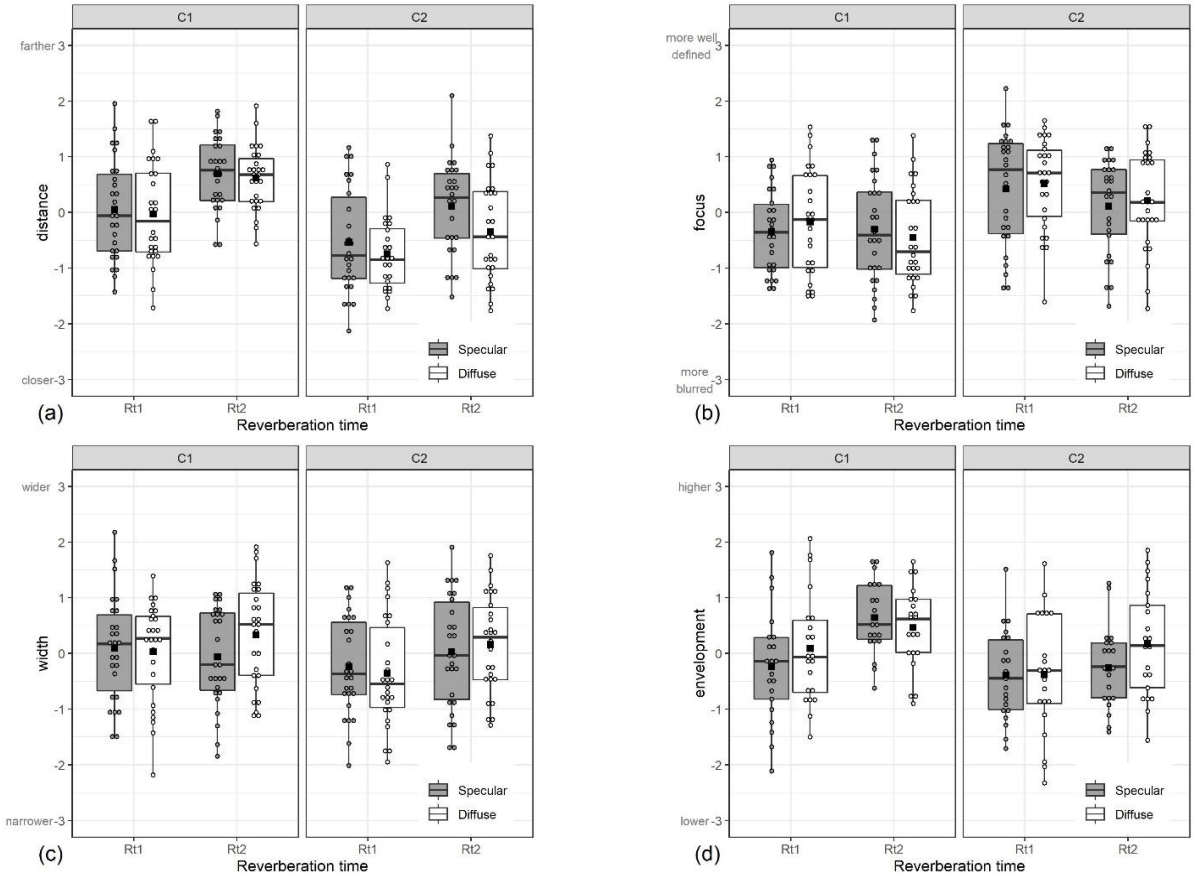

Figure 2 – Perceptual assessment (Z-scores) in quiet by type of reflection, reverberation time and clarity. Black squares are mean values and lines are the median values. Small circles are the single experimental values. a) *distance*; b) *focus*; c) *width*; d) *envelopment*.

For *focus* only the effect of C50 was significant [F(1,200)=26.71, $p<0.001$]; no other main effects nor interactions were found. Comparisons showed a more focused image as C50 was high (*C2 > C1*, $p<0.001$, *t.ratio*=5.17, *d*=0.717 – *medium/large*). The analysis of w*idth* did not show any effect for the three factors neither for interactions. Finally when *envelopment* was considered, Rt was a significant factor [F(1,159)=12.58, $p<0.001$] and also C50 [F(1,159)=11.08, $p=0.001$]. The type of reflection and interactions were not significant. Post-hoc analysis showed that Rt2 evoked a more enveloping sound source (*Rt2 > Rt1*, $p<0.001$, *t.ratio*=3.56, *d*=0.55 - *medium*) as also lower C50 did (*C1 > C2*, p=0.001, *t.ratio*=3.35, *d*=0.518 - *medium*).

4 Results of Experiment 2: spatial percepts in noisy conditions

The results for the assessment of spatial percepts in the noisy conditions are reported in Figure 3. For *distance* the effect of Rt was significant [F(1,238)=9.93, p=0.001], as that of U50 [F(1,238)=25.52, p<0.001] and also that of the type of reflection [F(1,238)=23.71, p<0.002]. No other significant effects were found across interaction. The post-hoc analysis revealed farther perception of the source for longer Rt (*Rt2 > Rt1*, p=0.001, *t.ratio*=3.15, *d*=0.401 - *small*), lower U50 (*U1 > U2*, p<0.001, *t.ratio*=5.06, *d*=0.645 - *medium*) or when a specular reflection occurred (*Specular > Diffuse*, p<0.001, *t.ratio*=4.87, *d*=0.621 - *medium*). Analyzing *focus* the significant factors were Rt [F(1,236)=4.41, p=0.037], U50 [F(1,236)=9.35, p=0.002], and type of reflection [F(1,236)=5.91, p=0.016], but there weren't significant interactions. Post-hoc comparisons revealed that the source was more focused with a short Rt (*Rt1 > Rt2*, p=0.038, *t.ratio*=2.09, *d*=0.268 - *small*), an higher U50 (*U2 > U1*, p=0.003, *t.ratio*=3.04, *d*=0.389 - *small*), and with a diffuse reflection (*Diffuse > Specular*, p=0.016, *t.ratio*=2.42, *d*=0.31 - *small*)

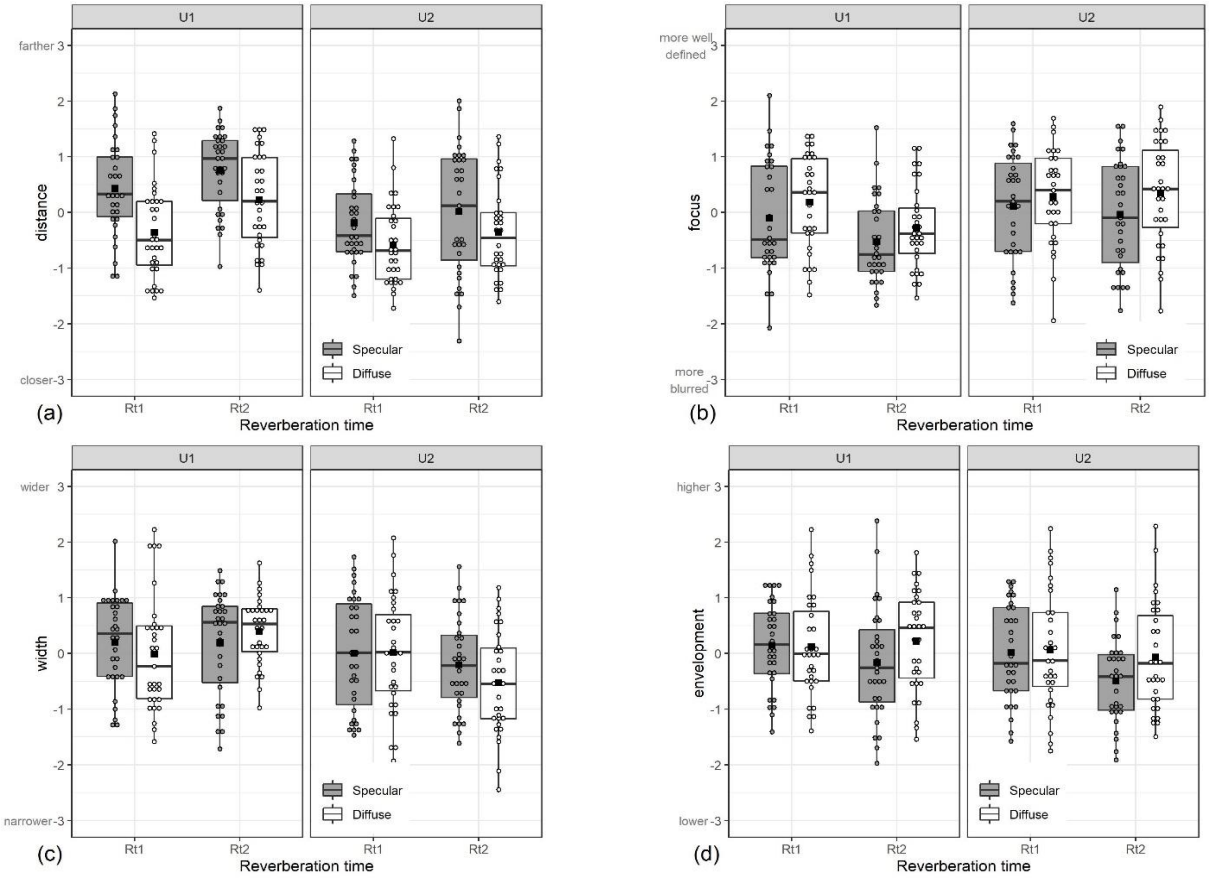

Figure 3 – Perceptual assessment (Z-scores) in noisy conditions by type of reflection, reverberation time and useful to detrimental ratio. Black squares are mean values and lines are the median values. Small circles are the single experimental values. a) *distance*; b) *focus*; c) *width*; d) *envelopment*.

The analysis of the *width* showed only the main effect of U50 [F(1,238)=10.13, p=0.002), but not that of either Rt (p=0.43) or type of reflection (p=0.49). The interaction between U50 and Rt was significant [F(1,238)=5.99, p=0.015], revealing a wider source perception for longer reverberation time when U50 was high too (U2: *Rt1 > Rt2*, p=0.046, *t.ratio*=2.29, *d*=0.411 - *small*) and also for lower U50 when Rt was longer (*Rt2: U1 > U2*, p<0.001, *t.ratio*=3.99, *d*=0.72 - *large*). As regards the *envelopment*, no significant effect was found either for the three main factors or for the interactions (all ps > 0.09).

5 Results of Experiment 3: speech intelligibility in noisy conditions

In Figure 4 the speech intelligibility data for the noisy conditions are shown. The effect of Rt was slightly significant [$\chi2(1)=3.83$, $p=0.05$], the effects of U50 [$\chi2(1)=37.17$, $p<0.001$] and type of reflection [$\chi2(1)=12.92$, $p<0.001$] were both significant. In addition the interaction between U50 and Rt was significant [$\chi2(1)=5.31$, $p=0.021$]. Post-hoc run on the type of reflection showed higher intelligibility with diffuse early reflections (*Diffuse* > *Specular*, $p<0.001$, *z.ratio*=3.57, *d*=0.169 - *very small*). Post-hoc comparisons for the interaction showed that for lower U50 one has higher SI if Rt is low (*U1*: *Rt1* > *Rt2*, $p=0.007$, *z.ratio*=2.93, *d*=0.21 - *small*), and that one has a higher SI for higher U50 independently of Rt (*Rt1*: *U2* > *U1*, $p=0.009$, *z.ratio*=2.70, *d*=0.182 - *very small*; *Rt2*: *U2* > *U1*, $p<0.001$, *z.ratio*=5.91, *d*=0.397 - *small*).

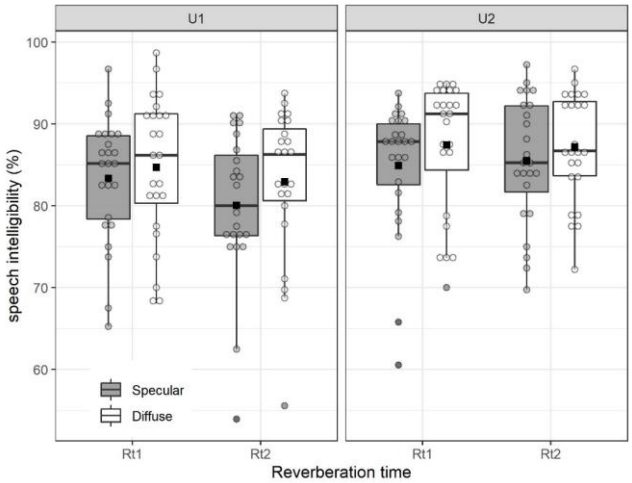

Figure 1 - Speech intelligibility in noisy conditions as a function of type of reflection, reverberation time and useful to detrimental ratio.

6 Discussion

6.1 Dependency of the spatial percepts on the acoustical variables

The first aim of the present experiment was to test the effect of clarity, reverberation and type of reflection on selected spatial percepts after manipulating HRIRs to obtain two clarity and two reverberation levels. The second aim was to investigate how the spatial percepts depend on the same variables in noisy conditions and whether the same percepts are distorted from quiet to noise. The third aim was to study if and how speech intelligibility in noise depended on the variables above. The ranges of the reverberation times and clarities, as well as the chosen SNR, are quite typical for mid-sized rooms for speech where acoustics has already been optimized. So, it was implicitly assumed that there were no specific faults in the acoustical characteristics. Furthermore, the presentation level was kept fixed to minimize level-related cues in the evaluations. As expected, in quiet *distance* was sensitive to the ratio of direct to early and late sound and to the overall amount of reverberation; interestingly the same trend was kept also with the presence of noise. But in the noisy conditions also the main effect of type of reflection showed up, with the specular cases being perceived as farther from the source. The evaluation of distance with noise has rather old, limited and contrasting evidence [13] due to the type of target signal, masking noise and experimental details; thus, it is unclear if noise increases or decreases the perceived source *distance*. *Envelopment* in quiet was partly explained by factors reverberation and clarity. When analysing the sound field in halls for music (not for speech, though) this quantity is traced back to the late (and possibly lateral [8]) energy reaching the listener; these are quantities derived from the sound strength G. If one assumes that a similar mechanism would be applicable

also to a speech source, the present results are consistent because the definition of the late sound strength G_L involves clarity. In other words, a lower clarity is paired with higher G_L and with a longer Rt too: it is thus easily understood how both trends cooperate in increasing *envelopment.* On the other hand, the C50 ad Rt main effects disappear with noise since the evaluations of the construct did not differ between conditions. This finding shows that information encoded into the late part of the HRIR cannot be entirely retrieved due to the masking of noise. The *width* in the quiet conditions did not show any dependence from clarity, reverberation and type of reflection. Although the judgement of the *width* for a speech source is not straightforward, past studies have demonstrated that listeners could discriminate *width* from *focus* for a speech source [11]. In the present case the variations of clarity were obtained by attenuating the reverberant tail while the early part of the HRIR was not touched. As a consequence, the changes in clarity implied a change in the overall HRIR energy and in the *relative* relationship between the early and the late sound, but practically only the late one changed in *absolute* terms (almost - 6 dB). Knowledge from concert halls tells us that *width* is usually predicted by the early sound (and by the lateral one in particular) [8] but can be affected also by the late sound. In fact, a change in *width* caused by a 1 dB change in the early sound can be obtained only with 6.8 dB change in the late sound [8]. So, it is quite plausible that the present change in the late sound was not enough to be detected as a change in *width*. Interestingly, in the noisy conditions an interaction between Rt and U50 showed up. This was unexpected as long as already in the quiet conditions it was not possible to differentiate *width.* In the higher U50 and shorter Rt the source was perceived wider, and in the longer Rt and lower U50 was still perceived wider. The former finding is consistent with the previous view based on the importance of the early lateral and upward-arriving [14] energy for *width*[1], while the latter finding is not. So, it is unclear how this unmasking of *width* in noise is realized, and specific experiments shall investigate the phenomenon in detail. In this respect it has also to be remarked that in the literature there is no study that tackled the changes of *width* of a speech source in noise. The spatial percept of *focus* was sensitive only to the clarity change in quiet so that a higher clarity was perceived as a more focused sound image. As for the *width*, the addition of noise made *focus* much more easily detected because the main effects of reverb and reflection type were reported. In these cases the sound source appears more focused for high clarity, shorter reverb and with diffuse reflections. The perceptual basis of *focus* are under investigation and a first qualitative hypothesis is that focus depends more on the nature of the early reflections, and not only on their direction and energy like *width* [12]. As a matter of fact, noise unveils some of the features that are necessary to decipher *focus* which were hindered by reverberation. The effect of reverberation itself becomes evident and this may be due to the gained ability to judge even the first part of the sound decay given that the later part is completely masked by noise. Strikingly, also the reflection type plays a role for *focus* in noisy conditions whereas this did not happen in quiet. In a previous work [12] HRIRs were used without reverberation ("dry" case) having only the three early reflections (either specular or diffuse) like in the present work. When the two "dry" conditions (three specular vs. three diffuse early reflections) were compared in that experiment it was shown that *focus* in quiet was significantly higher for the specular case. The present experiment adds that with reverberation, but still in quiet, the cues induced by the type of reflection are masked. When noise is added the type of reflection is again important but the effect goes in the opposite direction (diffuse>specular). To be better understood, it is necessary to disentangle the effect of noise and reverberation on *focus* by evaluating the percept in "dry" but noisy conditions. At present it will be demonstrated in par. 6.2 that this behaviour is justified by the fact that noise hampers the *focus* more for specular reflections than for diffuse ones.

6.2 Comparison of the spatial percepts in quiet and in noise

In order to directly compare the quiet and noisy conditions a set of t.test or Wilconxon tests (for not normal distributions) was accomplished. For this analysis the raw scores were considered because the two panels' z-score data were referred respectively to the quiet and the noisy conditions and could not be directly compared. Tab. 3 reports the results of the comparisons over the eight conditions, quiet (Q) vs noise (N). One can see that in none of the cases the sound image is either more or less distant, more focused, wider or

[1] The term "image size" is used in [14] rather than *width*; the latter is assumed here to include the former in the experiments because a specific evaluation of upward image spread was not pursued.

more enveloping in noise than in quiet. In the case of *distance* there are no significant differences and it has to be recalled that in both quiet and noise the sound levels were fixed (quiet: 57 dBA; noise 64 dBA i.e. SNR= - 6dB) so that the loudness cue was minimized. The overall loudness was increased in noise (7.5 sone) but with negligible discrepancies between conditions. In the *envelopment* only one condition shows Q > N, while in half of the conditions for *focus* and *width* one has that the source appears either more focused or wider, or both in one condition (last raw of Tab. 3).

Table 3 – Results of the t.test or Wilcoxon tests for the comparison of quiet (Q) and noisy (N) cases. The conditions are marked for reverberation (Rt1, Rt2), clarity or useful-to-detrimental ratio (C1, C2 for both in this case) and for the diffuse (D) or specular (S) early reflections. In bold the significant cases. Grey background for S significant cases in *focus*, and green background for D significant cases in *width*.

Condition	***Distance***	***Focus***	***Width***	***Envelopment***
Rt1_C1_D	p = 0.57	p = 0.72	***p* = 0.049**, t=2.01, Q > N	p = 0.80
Rt1_C1_S	p = 0.26	p = 0.38	p = 0.089	p = 0.66
Rt1_C2_D	p = 0.86	***p* = 0.017**, t=2.46, Q > N	p = 0.40	p = 0.42
Rt1_C2_S	p = 0.12	***p* = 0.007**, t=2.79, Q > N	p = 0.62	p = 0.66
Rt2_C1_D	p = 0.41	p = 0.21	***p* = 0.043**, W=513.5, Q >N	p = 0.56
Rt2_C1_S	p = 0.75	***p* = 0.009**, t=2.72, Q > N	p =0.17	***p* =0.017**, W=513.5,Q>N
Rt2_C2_D	p = 0.95	p = 0.28	***p* = 0.001**, W=513.5, Q >N	p = 0.22
Rt2_C2_S	p = 0.94	***p* = 0.047**, t=2.04, Q > N	***p* = 0.042**, t=2.08, Q > N	p = 0.10

In particular 3 out of 4 conditions for *focus* are related to specular reflections and 3 out of 4 conditions for *width* involve diffuse reflections. Although not complete, these findings allow to depict some features of the distortions that the sound image undergoes when noise is added. Typically, if early reflections are specular the *focus* is decreased (*source blurring*) while *width* is not much touched, whereas if early reflections are diffuse *width* is decreased (*source shrinking*) whereas the *focus* is not changed much. In the "dry" experiment of [12] recalled in par. 6.1 *width* did not depend on the reflection type, and the same happened in the present reverberant quiet and noisy conditions when they are analysed separately. Noise hampers preferable diffuse reflections in case of *width* but the *source shrinking* that occurred when comparing *width* in quiet and noise is thus not sufficient to highlight a difference in *width* between specular and diffuse reflections in noisy conditions.

6.3 Speech intelligibility and the sound image in noise

The speech intelligibility experiments were pursued only in noise because the quiet conditions would have prevented to record any significant difference from ceiling. The results showed how speech intelligibility is modulated by the amount of early reflections which are integrated into the direct sound [5] in particular up to the time limit of 50 ms [1]. Moreover, Rt was also a significant factor; its interaction with U50 shows that higher clarity in both reverberations, and shorter reverberation ensure better SI. All of this knowledge was expected and adheres to the notion that our auditory system is capable of aggregating favorable contributions while it is teased by longer sound tails even at fixed clarity values. But in the present experiment there was a further manipulation that could not be traced either by U50 or Rt, that is the type of reflection. In fact, this variable provided undistinguishable U50 and Rt data (see Tab. 2) because timing, amplitude and directions of the two sets of reflections were the same. On the other hand, the binaural information was altered because the phase relationships of diffuse rather than specular reflections were different at the ears. The SI results favored the diffuse case. From the previous argument this finding appears to be a binaural effect which is not tied to the ability of the auditory system to integrate early energy. Moreover, the analysis of spatial percepts in quiet and in noise allowed to depict the sound source in both conditions and to compare how this picture was affected by noise. In particular it is possible to drive an *association* between the present speech intelligibility outputs and the description of the auditory image of the sound source. Compared to quiet, in noise the sound source appears at unvaried distance, but it is perceived in most cases as less focused with specular reflections, and less wide for diffuse ones. In both quiet and noise cases the dependency of the

spatial percepts was analyzed separately and, interestingly, some of the acoustical indicators that were not significant in quiet became important in noise. This means that when forming the speaker auditory image, our auditory system in quiet seems not to make complete use of (or is not able to entirely parse) the fine details of the early sound while in noise, being the later arriving contributions masked by noise and the earlier consequently unmasked, it is able to grasp relevant details and to use them in order to extract the signal from the noise. In particular this is the case for the type of reflection; *focus* and *distance* in noise depended from the three variables as main factors whereas the type of reflection was not a factor in quiet. In noise the diffuse reflections provide always a closer and more focused sound image compared to specular ones: both cues of *distance* and *focus* can be hypothesized to provide a perceptual unmasking of the target signal from diffuse noise that ensures a better SI. On the other hand, the dependence of *width* on the acoustical variables in noisy conditions is less systematic and does not involve the type of reflection, so its contribution does not seem as crucial (i.e. the type of reflection is a factor for SI but not for *width*). It can be remarked that the behavior of *width* is also consistent with [10], where the distinction with *focus* was not considered. Finally, *envelopment* is not sensitive to any acoustical variable or to the reflection type in noisy conditions so its involvement into the formation of sound image that backs the SI performance is regarded of a lesser importance.

References

[1] Bradley, J. S., Sato, H., & Picard, M. On the importance of early reflections for speech in rooms. *J Acoust Soc Am*, 113(6), 2003, 3233-3244.

[2] Soulodre, G. A., Popplewell, N., & Bradley, J. S. Combined effects of early reflections and background noise on speech intelligibility. *J Sound Vib*, *135*(1), 1989, 123-133.

[3] Arweiler, I., & Buchholz, J. M. The influence of spectral characteristics of early reflections on speech intelligibility. *J Acoust Soc Am*, *130*(2), 2011, 996-1005.

[4] Warzybok, A., Rennies, J., Brand, T., Doclo, S., & Kollmeier, B. Effects of spatial and temporal integration of a single early reflection on speech intelligibility. *J Acoust Soc Am*, 133(1), 2013, 269-282.

[5] Rennies J, Warzybok A, Brand T, Kollmeier B. Measurement and Prediction of Binaural-Temporal Integration of Speech Reflections. *Trends Hear*, 2019, 23:2331216519854267.

[6] Kohlrausch, A. G., Braasch, J., Kolossa, D., & Blauert, J. An introduction to binaural processing. In J. Blauert (Ed.), *The Technology of Binaural Listening* (pp. 1-32). Berlin, Heidelberg: 2013. Springer.

[7] Kolarik, A. J., Moore, B. C et al. Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss. *Atten Percept Psycho*, *78*(2), 2016, 373-395.

[8] Bradley JS, Reich RD, Norcross SG. On the combined effects of early- and late-arriving sound on spatial impression in concert halls. *J Acoust Soc Am*, 108(2), 2000, 651-61.

[9] Cubick J, Buchholz JM, et al. Listening through hearing aids affects spatial perception and speech intelligibility in normal-hearing listeners. *J Acoust Soc Am*, 144(5), 2018, 2896-2905.

[10] Ahrens A, M Marschall M, Dau T. , The effect of spatial energy spread on sound image size and speech intelligibility, *J Acoust Soc Am*, 147(3), 2020, 1368-1378.

[11] Visentin C, Pellegatti M, Prodi N, Effect of a single lateral diffuse reflection on spatial percepts and speech intelligibility, *J Acoust Soc Am*, 148 (1), 2020,122-140

[12] Visentin C, Pellegatti M, Prodi N, Effects of multiple early diffuse reflections on spatial percepts, in revision for *J Acoust Soc Am*, 2021.

[13] Cabrera D, Gilfillan D, Auditory distance perception of speech in the presence of noise, Proc. of Int Conf. on Auditory Display, Kyoto, Japan, July 2002.

[14] Furuya H, Fujimoto K, Takeshima Y, Nakamura H, Effect of early reflections from upside on auditory envelopment, J. Acoust. Soc. Jpn. (E) 16(2), 1995, 97 – 104.