ADJUSTING THE PERCEIVED DISTANCE OF VIRTUAL SPEECH SOURCES BY MODIFYING BINAURAL ROOM IMPULSE RESPONSES

Robert Albrecht

Aalto University
Department of Media Technology
P.O. Box 15500, FI-00076 Aalto, Finland
robert.albrecht@aalto.fi

Tapio Lokki

Aalto University
Department of Media Technology
P.O. Box 15500, FI-00076 Aalto, Finland
tapio.lokki@aalto.fi

ABSTRACT

Effective control of the perceived location of virtual sound sources is an important aspect of auditory displays. While room acoustics modelling may be used to produce cues related to the sound source and listener location in a space, in many real-time applications it is more feasible to utilize ready-made room impulse responses. This paper looks at how the perception of distance can be affected by modifying the temporal envelopes of room impulse responses. Two measured binaural room impulse responses were modified by amplifying or attenuating different portions of them before convolving them with speech samples. Listeners were asked to judge the relative distances between these virtual speech sources presented over headphones. The results suggest that the perception of distance is more effectively altered by modifying an early-to-late energy ratio, where approximately 50–100 ms of the impulse response is included in the early energy, than by directly modifying the traditional direct-to-reverberant energy ratio.

1. INTRODUCTION

Auditory distance perception has been extensively studied to identify different factors affecting the perceived distance of sound sources. These factors include sound intensity, spectrum, and binaural cues [1]. In a reverberant environment, one possible cue is also the direct-to-reverberant energy ratio. Inside an enclosed space, late reverberation can be approximated by a diffuse sound field, and the reverberant energy is independent of sound source distance. On the other hand, the pressure of sound arriving directly from a point source is proportional to the distance travelled. Thus, according to this approximation, the direct-to-reverberant energy ratio decreases as the distance to the sound source increases. This relationship is widely accepted as being used by humans as an important cue in auditory distance perception [1, 2].

When calculating the direct-to-reverberant energy ratio of room impulse responses, typically a 2–3-ms time window is utilized to determine the direct energy portion of the ratio [3, 4]. All the energy outside this window, including that of early reflections, is considered reverberant. An increase in early reflection energy lowers this ratio and should thus result in an increase in perceived distance. On the other hand, it is recognized that early reflections integrate with the direct sound and make it louder [5]. This suggests that added early reflection energy might actually decrease the perceived distance of a sound source. If this is the case, it implies that the direct-to-reverberant energy ratio is not the most appropriate measure of the ways in which the temporal envelope of room impulse responses affects distance perception. Instead, an early-to-late energy ratio might correlate more strongly with the perceived distance.

Previously, a modified direct-to-reverberant energy ratio has been suggested by Bronkhorst and Houtgast, using a 6-ms time window for calculating the direct energy [6]. With this modified ratio, the perceived distance of virtual sound sources could be predicted in two experiments utilizing modelled room acoustics. This paper, however, takes a different approach and modifies the temporal envelopes of measured binaural room impulse responses. This allows to examine the effects that amplification or attenuation of different time segments of room impulse responses have on the perceived distance of virtual sound sources.

The research leading to these results has received funding from Nokia Research Center, the Academy of Finland, project no. [257099], and the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no. [203636].

2. AUDITORY DISTANCE CUES

Intensity Assuming a point source in an acoustic free field, the sound pressure level drops by 6 dB as the distance from the source doubles. In case the sound source is familiar, listeners can take advantage of this fact when estimating the distance to the source. However, studies have shown that a change in sound pressure level greater than 6 dB is required to change the perceived distance by a factor of two [7].

Familiarity with a sound source might also provide misleading cues. In experiments with live speech sources, blindfolded listeners have been shown to underestimate the distance of whispered voices and overestimate that of shouted voices [8]. However, if this effect is taken into account, it could be used to effectively control the location of virtual speech sources over a large range of distances [9].

Direct-to-reverberant energy ratio The direct-to-reverberant energy ratio (D/R) is calculated from the impulse response $h(t)$ between two locations using the equation

$$D/R = \frac{\int_0^T h^2(t)dt}{\int_T^\infty h^2(t)dt}, \quad (1)$$

where $T \approx 2 - 3ms$, separating the direct sound from all reflections.

To explain the results of two experiments, Bronkhorst and Houtgast proposed a modified direct-to-reverberant energy ratio as a cue for auditory distance perception, where an integration window of 6 ms was used to calculated the energy of the direct sound [6]. The experiments were performed using artificial binaural room impulse responses with a maximum duration of 0.11 s together with bursts of pink noise. The suggested 6-ms integration window is long enough to include the energy of some of the earliest reflections with the energy of the direct sound.

Bronkhorst later suggested a binaural model for auditory distance perception, where the separation of direct and reverberant energy was performed spatially rather than temporally [10]. An interaural-time-difference (ITD) window was used to determine which energy was considered direct and which was considered reverberant. This new model was able to explain the results of the original experiments [6] as well as new experiments where the temporal model made poor predictions of the results.

Larsen et al., on the other hand, found in their experiments that D/R discrimination is not based on binaural cues such as the interaural cross-correlation (IACC) [4]. In these experiments, monaurally obtained just-noticeable differences (JNDs) of D/R did not differ significantly from binaurally obtained. However, Larsen et al. pointed out that these D/R discrimination results cannot directly be linked to auditory distance perception tasks.

Other cues As a sound source approaches the listener's head at distances smaller than approximately 1 m and from a direction away from the median plane, the interaural level difference (ILD) increases significantly due to an increased head-shadowing effect as well as an increased difference in distance attenuation [11]. The ITD, however, is almost independent of distance. Listeners are able to utilize these binaural cues when judging the distance of nearby sources [12]. These cues could thus be useful for constructing near-field virtual auditory displays.

Auditory localization has been shown to be affected by visual cues under certain conditions [1]. This is the case when it comes to both direction and distance. An appropriate visual target may cause the auditory event to be pulled towards its location. Visual information may also aid in creating a representation of the environment in memory that improves the perception of distance after visual cues have been removed [13].

Absorption of sound in air will cause high frequencies to attenuate more than low frequencies. Low-pass filtering of signals has been shown to increase the perceived distance of the sound source, but such noticeable effects are typically caused by the attenuation of high frequencies when sound reflects off surfaces rather than by air absorption [4]. Larsen et al. suggest that changes in the direct-to-reverberant energy ratio are discriminated mainly based on these spectral cues [4].

Combination of cues Experiments by Mershon and King suggest that reverberation is an absolute cue in auditory distance perception, while intensity serves only as a relative cue [14]. These experiments were, however, performed using noise bursts. For familiar sounds, such as human speech, listeners have through experience developed a mapping between intensity and the distance of the sound source. It has been shown that listeners can make reasonably accurate estimates of the distance of live talkers in anechoic conditions [9].

Zahorik has presented a hypothetical framework for the combination of different cues to auditory distance [15]. He suggests that the role of each cue in producing the final distance percept depends highly on their availability and quality in the current scene. Cues are not only weighted less if they are weak or unavailable, but also if they conflict heavily with the other cues available. In Zahorik's experiments, the intensity cue dominated over the direct-to-reverberant energy ratio when speech stimuli were used. With noise-burst stimuli, however, both cues were weighted approximately equally.

3. METHODS

To obtain knowledge about the effects that modifications of the temporal envelope of binaural room impulse responses (BRIRs) have on auditory distance perception, a listening test was devised. Two BRIRs, measured with a dummy head, were used, both taken from the Aachen Impulse Response Database [16, 17]. The first impulse response was measured in a moderately reverberant stairway (space I). The second impulse response was from an aula originally built as a church (space II), thus being an example of a highly reverberant environment. The energy envelopes of the two impulse responses are shown in Fig. 1.

One criterion for choosing the mentioned spaces from the impulse response database was that impulse responses had been measured at different azimuths, including 90°, only in these two spaces. Informal listening prior to the listening tests revealed that the impulse responses measured at small azimuthal angles provided poor externalization. This phenomenon has previously been reported by, e.g., Begault [18]. For this reason, only impulse responses measured at 90° azimuth were selected for this study. Another reason to exclude source locations inside the central field of view was to avoid clear visual stimuli that could affect the auditory distance perception. The impulse responses were measured at an elevation angle of 0°, typical for everyday speech sources.

From the different distances available in the database, impulse responses measured at a distance of 3 m from the source were selected. This distance represents a suitable starting point for modification of the distance perception, allowing closer and farther distances common for sound sources indoors. Additionally, binaural cues to distance perception are practically nonexistent at this distance [1].

As sound samples, the sentence "we talked of the sideshow in the circus," spoken by a female voice and a male voice, was used (samples "FB07_01" and "MB07_01" from the TSP speech database [19]). The samples were recorded in an anechoic chamber. The main reason for using speech samples in this study was the applicability of the results in communication applications.

Two different sets of modifications were done to the BRIRs before they were convolved with the speech samples. In modification set A, a portion of the impulse response starting at a specific point in time after the direct sound and continuing until the end of the response, was amplified or attenuated by 6 dB. In modification set B, a portion of the impulse response beginning directly after the direct sound and ending at a specific point in time, was amplified or attenuated by 6 dB. The modifications are illustrated in Fig. 2. Both modification sets also included the unmodified BRIR as well as one version where the whole impulse response was amplified 6 dB and one version where the whole impulse response was attenuated 6 dB. The combination of two different spaces, two different speech signals, and two modification sets resulted in a total of eight test cases, with thirteen samples each.

The test cases were presented to participants in random order.

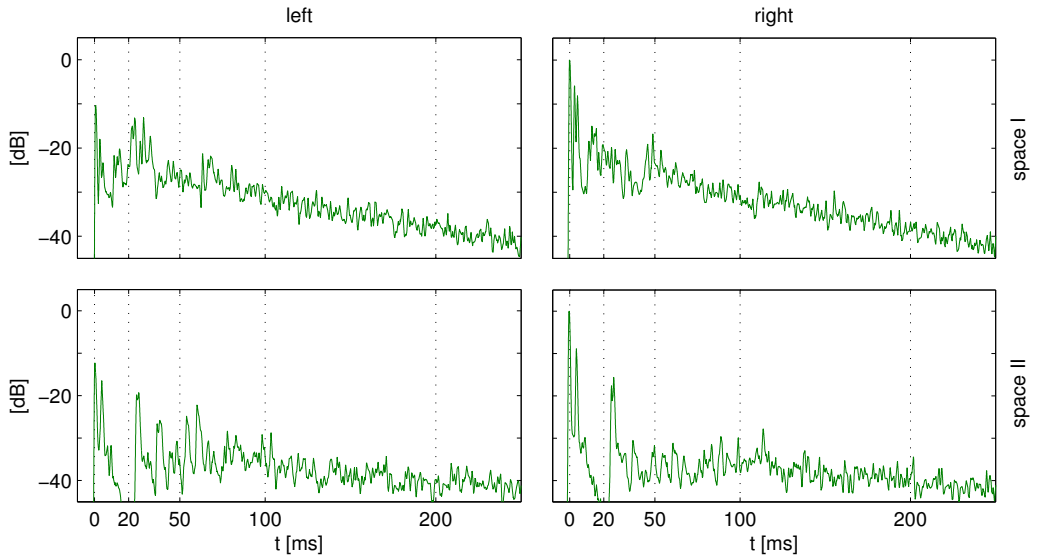


Figure 1: Energy envelopes of the binaural room impulse responses.

A condensed version of the user interface for the main listening test is shown in Fig. 3. Each of the thirteen samples was represented by a play button, with which the sample was played. The unmodified reference sample had a fixed position in the middle of the interface, but all the other samples could be dragged and dropped to different positions. Participants were asked to place each sample on the horizontal axis based on how distant the sound source in that sample sounded. Samples that sounded the closest were placed to the left and samples that sounded distant towards the right end of the axis. Apart from the reference sample, there were no fixed positions on the axis. Thus, there was no fixed point corresponding to the listener's head. In addition to ordering the samples, from the closest to the most distant, participants were asked to take into account the relative differences in the perceived distance of the samples. Samples that seemed to be at almost the same distance should thus be grouped closely together, while samples that seemed to be at clearly different distances should be spaced apart.

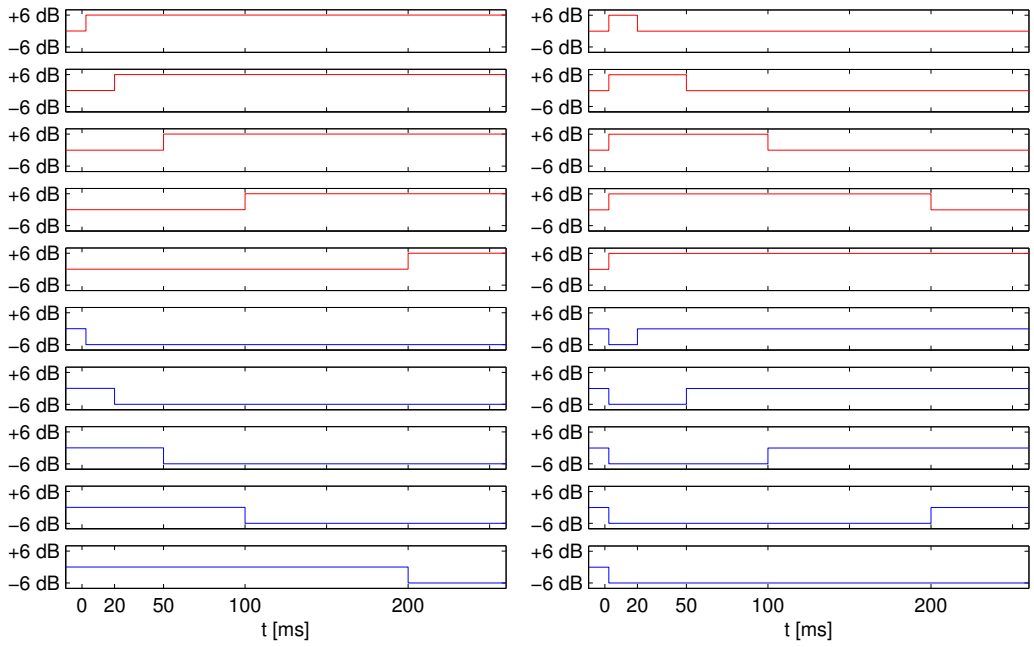
Figure 3: The interface used in the first listening test. Sound samples can be played and dragged around, so that their placement on the horizontal axis in the end corresponds to the distance of the virtual sound source in the sample. The interface in the figure is a condensed representation of the actual interface.

Participants also did a second listening test straight after the first one. Listeners were asked to specify, in meters, how far away the virtual sound source in each sample sounded. For this test, only the two unmodified BRIRs were used, plus versions of these where the whole impulse response was either amplified or attenuated 6 dB. This test utilized the same male and female speech samples as the first test and included two repetitions of each combination of impulse response and speech sample. The order of the resulting 24 samples was randomized, and the samples were presented one at a time. Participants could listen to a sample as many times as they wanted. This test was mainly performed to give an absolute distance reference for some of the samples used in the first listening test, as only relative distances were judged in that test.

Sennheiser HD 650 headphones were used in the listening tests. A comfortable listening level was chosen by the experimenters prior to the tests, and listeners were not allowed to adjust this level. Interestingly, as is shown in Sec. 4.4, the choice of level resulted in participants perceiving the unmodified reference samples to be at a median distance of 3–3.5 m. This corresponds with the 3-m distance used in the BRIR measurements. The tests were performed in two quiet and small office rooms.

4. RESULTS

Of the 24 people who participated in the listening tests, 22 worked at the university and two worked in the industry. The participants, of which 4 were female and 20 male, represented ages between 26 and 42 years. Exactly half of the participants worked with tasks related to audio signal processing and acoustics, while the other half did not have experience in the area. Two of the participants reported having minor hearing impairments, while the rest were not aware of any impairments.



(a) Modification set A. (b) Modification set B.

Figure 2: Modifications made to the binaural room impulse responses. The maximum amplitude of the direct sound is at $t = 0ms$ (averaged over both channels). Amplifications and attenuations beginning instantly after the direct sound start at $t = 2.4ms$.

4.1. Relative distance estimates in space I

Because participants were allowed to use the answering scale in the first listening test freely, apart from the reference sample having a fixed position, Z-scores of each participant's answers to each case were calculated separately. This compensates, to some extent, for the participants' different use of the scale. Furthermore, the Z-scores were divided by 6 and 0.5 was added to the result, giving positive values between 0 and 1. These modified Z-scores are used in the analysis and presentation of the distance estimates.

Fig. 4 shows the medians of the estimated distances in space I, together with their 95% bootstrap confidence intervals (calculated using the *bootci* function in MATLAB with 10 000 samples). Bootstrapping was chosen because the distributions of many of the distance estimates could not be approximated well by any common distribution.

Of all the impulse response amplifications in modification set A, amplifying the impulse response from 50–100 ms onwards results in the largest increase in the perceived distance (see Fig. 4(a) and 4(b)). Similarly, attenuating the impulse response from 50–100 ms onwards results in the largest decrease in the perceived distance.

Controlling the direct-to-reverberant energy ratio by amplifying or attenuating the whole impulse response efter the direct sound does not result in the desired behaviour, but rather the opposite. A reduced D/R here results in a reduction of the perceived distance, while an increased D/R results in a slight or no increase in the perceived distance compared with the reference sample. This behaviour can at least partially be attributed to the fact that the attenuations or amplifications performed affect the loudness of the sample in a way that counteracts the D/R modification. However, this also raises the question, whether the direct-to-reverberant energy ratio correlates very well with the perceived distance of a sound source, or if another type of early-to-late energy ratio might correlate better with the perceived distance.

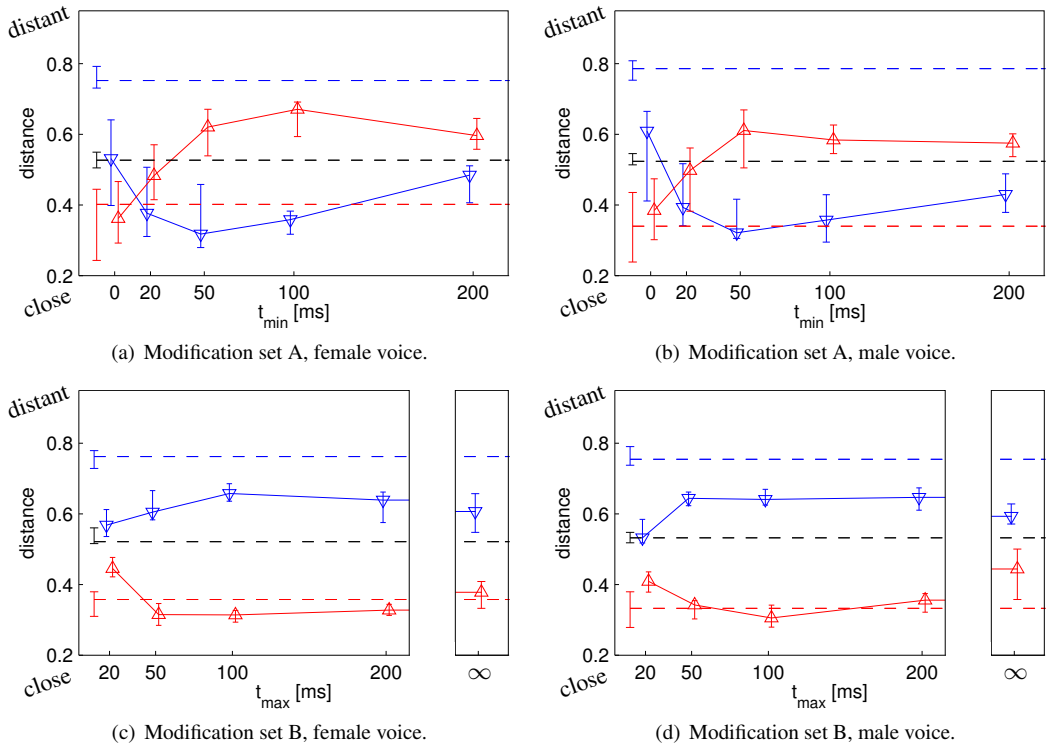
It should be pointed out that conventionally when modifying the D/R, it is the direct sound that is amplified or attenuated, and not the later part of the impulse response. However, it is clear from Fig. 4(a) and 4(b) that amplifying or attenuating the first 50–100 ms of the impulse response instead of only the direct sound will have a considerably larger effect on the perceived distance.

From the distance estimates of modification set B (Fig. 4(c) and 4(d)) it can be seen that amplifications of the impulse response starting after the direct sound and including up to 100 ms reduce the perceived distance. Similarly, attenuating this portion of the impulse response increases the perceived distance. When the portion of the impulse response included in the attenuation or amplification extends to 200 ms or more, an effect in the opposite direction is introduced. As with modification set A, it is unclear exactly how large an effect the increase or decrease of the loudness associated with each modification has on the perceived distance.

4.2. The effect of intensity and the early-to-late energy ratio

To look closer at the effect that the early-to-late energy ratio and intensity have on the perceived distance, a model taking these two factors into account was fitted to the distance estimates. The early-to-late energy ratio is calculated using the equation

$$C_T = \frac{\int_0^T h^2(t)dt}{\int_T^\infty h^2(t)dt}, \quad (2)$$



(a) Modification set A, female voice.

(b) Modification set A, male voice.

(c) Modification set B, female voice.

(d) Modification set B, male voice.

Figure 4: Distance estimates in space I: medians and their 95% confidence intervals. The dashed line in the middle represents the reference sample. The lower dashed line represents the 6 dB amplification of the whole sample, while the upper dashed line represents the 6 dB attenuation of the whole sample. The upward-pointing triangles represent the amplifications and the downward-pointing triangles the attenuations depicted in Fig. 2. For modification set A, these are located on the horizontal axis at the time when the attenuation or amplification of the impulse response begins. For modification set B, the placement on the horizontal axis corresponds to the time when the attenuation or amplification ends. The triangles are slightly offset from their actual positions on the horizontal axis to avoid overlapping.

where $h(t)$ is the impulse response and T is the time separating the early and the late part of the room impulse response. C_{50}, where T is chosen to be 50 ms, is often used as a measure of clarity for speech [20]. C_{80} is commonly used as a measure of clarity for music.

As a measure of the intensity of the stimuli, the equivalent continuous sound level was chosen as a basis:

$$L_{eq} = 10log_{10}\left(\frac{1}{\tau}\int_0^{\tau}\frac{x^2(t)}{x_{ref}^2(t)}dt\right)dB \tag{3}$$

Here, τ is the length of the measured signal $x(t)$ and $x_{ref}(t)$ is a reference signal. The linear L_{eq} has been shown to correlate fairly well with the perceived loudness of music and speech material [21]. The corresponding measure used in the distance model is the time-averaged energy:

$$\bar{E} = \frac{1}{\tau}\int_0^{\tau} x^2(t)dt \tag{4}$$

In the calculation of both C_T and $\bar{E}$, binaural loudness summation was taken into account, combining the level at the left and the right ears to produce a single measure

$$L_{mon} = g \cdot log_2(2^{L_{left}/g} + 2^{L_{right}/g}), \tag{5}$$

where the binaural gain g was chosen to be 3 dB [22].

The effect of C_T and $\bar{E}$ was taken into account both separately and combined, producing the distance estimates d_C, d_E, and d by means of the equations

$$d_C = a \cdot \left(\frac{1}{C_T}\right)^k, \tag{6}$$

$$d_E = b \cdot \left(\frac{1}{\bar{E}}\right)^l, \tag{7}$$

and

$$d = c \cdot \left(\frac{1}{C_T}\right)^k \cdot \left(\frac{1}{\bar{E}}\right)^l, \tag{8}$$

where a, b, c, k, and l are constants. A power function, inspired by Stevens' power law [23], is here applied to the physical measures C_T and $\bar{E}$ in an attempt to translate them into perceptual measures.

Fig. 5 shows the root-mean-square (RMS) error when fitting (8), (6), and (7) to the distance estimates in space I. This least-squares curve fitting was done with a trust-region reflective algorithm (using the *lsqcurvefit* function in MATLAB). The fitting was performed separately for all four cases and for (8), (6), and (7), to find the values of a, b, c, k, and l producing the smallest RMS error between the fitted distances and the distance estimates of all participants in each case. The reason for separating all cases was

that participants presumably used the answering scale differently depending on the case.

C_T and $\bar{E}$ were calculated for each listening test sample, i.e., from the modified BRIR in the case of C_T and from the combination of modified BRIR and speech sample in the case of $\bar{E}$. In addition, C_T was calculated separately for all values of T between 1 and 500 ms, with a 1-ms step size, and the fitting was done separately for all these values of T. To take into account the spectra of the speech samples, the BRIRs used to calculate C_T were first filtered according to the spectrum of the corresponding speech sample. Thus, frequencies not present in the listening test samples were also excluded when calculating this measure.

For modification set A (Fig. 5(a) and 5(b)), d_C (6) and d_E (7) give approximately the same error when fitted separately to the distance estimates. Combined (8), the error is reduced. Interestingly, the choice of modifications in modification set B (Fig. 5(c) and 5(d)), results in a considerably better fit for d_E than for d_C. Both combined, the error is only slightly reduced from that of d_E alone. Thus, it would seem that distance estimates for modification set B were made predominantly based on intensity.

The error when fitting (8) for modification set A is reduced when including up to 75–150 ms in the early part of the early-to-late energy ratio. The same tendency can be seen for modification set B, but the reduction is not very substantial in this case. When T is increased above 150 ms and up to 200 ms, the error increases. The error remains at a largely constant level when T is increased above 200 ms, where no step changes in the impulse response envelope where done in any of the modifications (see Fig. 2).

To illustrate the fitted distances versus the perceived distances, Fig. 6 displays the best fit of (8) compared with the medians of the distance estimates in space I when using the female speech sample and BRIR modification set A. For the calculation of C_T, T was chosen to be 100 ms, which is close to the optimal value as illustrated in Fig. 5(a). The RMS error of the fitted distances is 0.039 relative to the medians and 0.119 relative to the individual answers of all the participants. The optimum values for c, k, and l were in this case 0.088, 0.247, and 0.246, respectively.

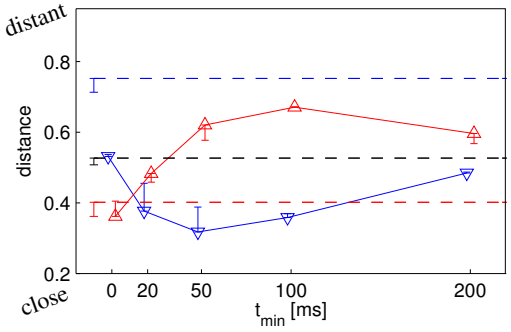


Figure 6: The error when fitting Eq. (8) to the distance estimates of space I, modification set A, female voice. Medians of the distance estimates are diplayed as in Fig. 4(a) and the error bars show the errors of the fitted distances relative to the medians.

4.3. Distance estimates in space II

The medians of the distance estimates in space II, together with their confidence intervals, are illustrated in Fig. 7. The impact of the different modifications is not as clearly interpretable as for space I (Fig. 4), but using a split point of 50–100 ms between the early and late energy would seem like a good starting point when modifying the perceived distance also in this case. However, especially the attenuations of the male voice in modification set A (Fig. 7(b)) show some interesting differences when compared with the same modifications in space I (Fig. 4(b)).

Fig. 8 shows the root-mean-square error when fitting d_C (6), d_E (7), and d (8) to the distance estimates in space II. As in space I, fitting d_C and d_E separately produces approximately the same error for modification set A, but the error of d_E alone is substantially smaller for modification set B. When fitting d and thus taking both effects into account, the error is reduced, but only considerably so for modification set A. For both modification sets, however, the error depends only little on the split point T between early and late energy.

4.4. Absolute distance estimates

The estimated absolute distances of the second listening test are illustrated in Fig. 9. The medians for the reference samples range from 3 to 3.5 m, while the medians for the +6 dB samples all are 2 m. The estimates for the -6 dB samples show larger differences between the two spaces, with the medians being 5 m for space I and 7.5 to 7.75 m for space II. For both spaces, the change in estimated distance when attenuating or amplifying the samples by 6 dB is markedly close to the ratio of one to two which would theoretically be expected in an acoustic free field.

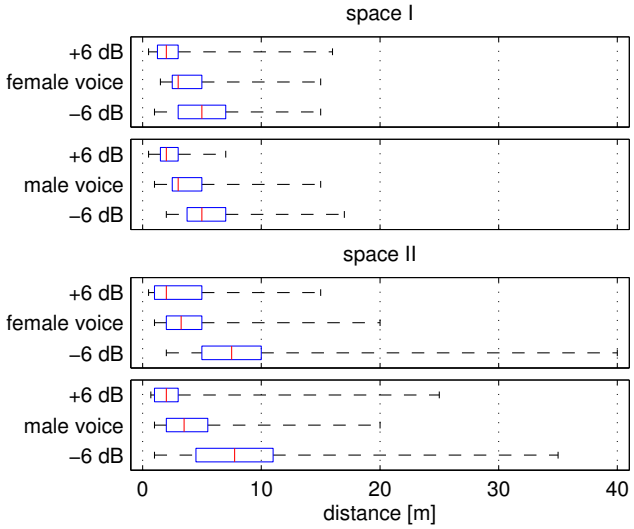
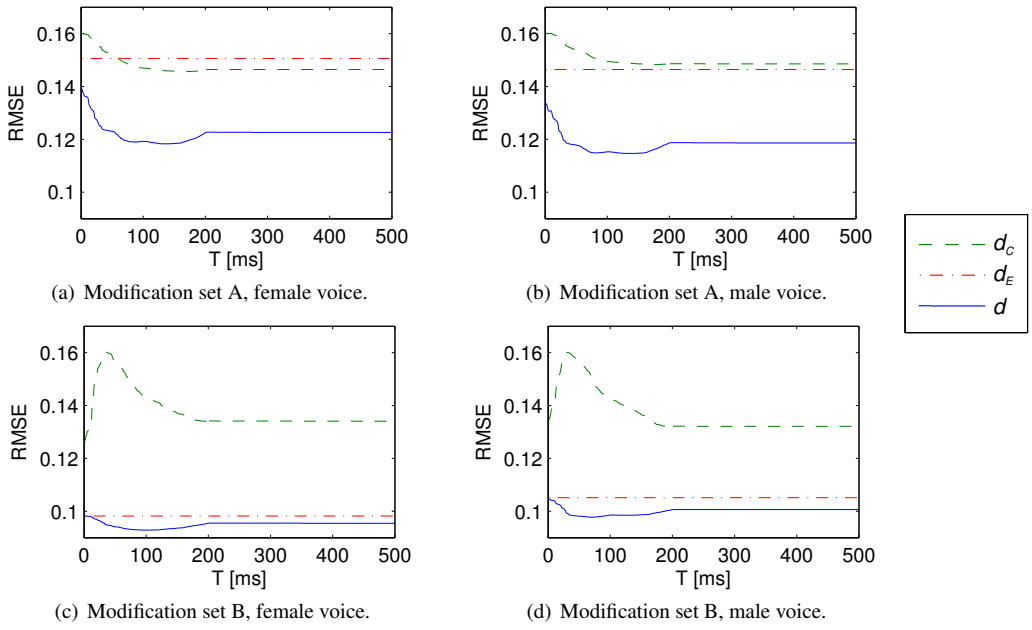


Figure 9: Box plots of the estimated absolute distances in the second listening test. The ends of the whiskers show the extreme values.

5. DISCUSSION

Participants were not explicitly asked how well the sound samples were externalized, but several participants said that the first sample they heard sounded like it came from a loudspeaker. Some even turned their head to the right to be sure this was not the case.



(a) Modification set A, female voice.

(b) Modification set A, male voice.

(c) Modification set B, female voice.

(d) Modification set B, male voice.

Figure 5: Root-mean-square error (RMSE) when fitting Eq. (6) (dashed line), Eq. (7) (dash-dot line), and Eq. (8) (solid line) to the distance estimates in space I, shown for different values of T.

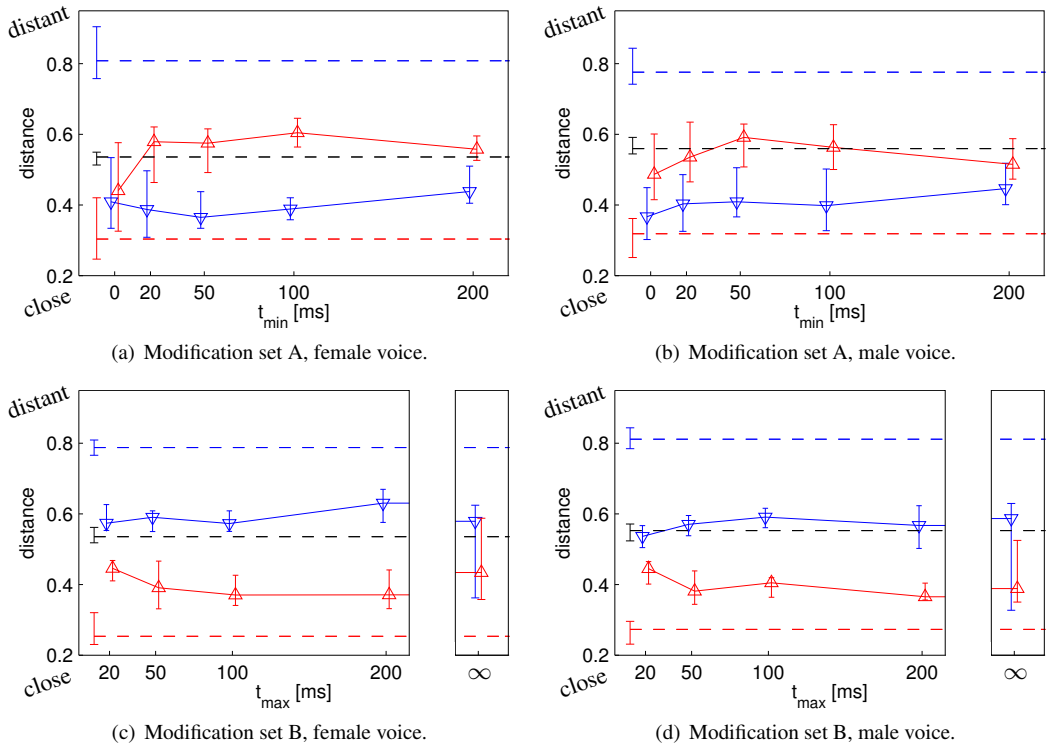
A few participants reported that some sound samples sounded unnaturally loud or amplified. Although the listening rooms used were small in size, many participants gave distance estimates in the second listening test far larger than the greatest dimension of the room. Only one participant commented that it was difficult to imagine a source at a greater distance than was possible inside the room. Most participants thought that the task was rather demanding. Including instruction and training, the test normally took between half an hour and one hour to complete.

To find different answering tendencies between participants, k-means clustering was performed based on the squared Euclidean distance between the distance estimates of the first listening test. Dividing the participants into two clusters produced one group of 15 participants and one group of 9 participants. Of the 12 participants working with tasks related to acoustics and audio signal processing, 7 were in the smaller group. When fitting (6), (7), and (8) separately for both groups, the main difference which could be observed was that the larger group apparently put a stronger emphasis on loudness and a weaker emphasis on the early-to-late energy ratio when making the distance judgements than all participants as a whole did. Correspondingly, the smaller group seemed to put less emphasis on loudness and more on the early-to-late energy ratio. Dividing the participants into three groups kept the larger group intact and split the smaller group into one group of 7 participants (of which 5 were audio professionals) and one group of 2 participants (both audio professionals), so no analysis of further clustering was deemed necessary.

In these listening tests, participants were asked how distant each virtual sound source sounded. How well the answers correspond with the distance of the auditory event is unclear. In the second listening test, where answers were given in meters, it is possible that participants tried to estimate the likely distance of the source rather than just saying at which distance the auditory event occured. However, in the first test, no absolute distance scale was used, and participants compared samples with each other and ordered them based on this comparison. It is more likely that participants compared the actual distances of the auditory events here.

For space I, the results clearly show that amplifying or attenuating the first 50–100 ms of the room impulse response has a larger impact on the perceived distance of the speech sources than does amplifying or attenuating the direct sound alone. However, no clear conclusions can be drawn from the results for space II. Looking at the impulse responses in Fig. 1, one could speculate that both the temporal distribution and the energy of the early reflections might explain the difference between the two spaces. Whereas the early reflections are temporally densely spaced in space I, the spacing is much sparser in space II, with a 20-ms gap between the floor reflection and the following reflection. The energy of the early reflections, compared with the direct sound, is considerably larger in space I than in space II. It could thus be hypothesized, that the type of modifications performed in this study are most effective when applied to BRIRs from relatively small spaces, where there is more early energy to modify. The spatial distribution of the early reflections might also have an impact, but this aspect is not looked into in the current study.

Although the results suggest that an early-to-late energy ratio is a useful measure when modelling and modifying auditory distance perception, it should be pointed out that these experiments do not tell us about the exact mechanisms by which the human auditory system interprets reverberation cues to produce auditory distance percepts. Although a temporal model was fitted to the perceived distances, this does not rule out that the mechanism is actually based on spatial, spectral or possibly other properties.



(a) Modification set A, female voice.

(b) Modification set A, male voice.

(c) Modification set B, female voice.

(d) Modification set B, male voice.

Figure 7: Distance estimates in space II: medians and their 95% confidence intervals. The dashed line in the middle represents the reference sample. The lower dashed line represents the 6 dB amplification of the whole sample, while the upper dashed line represents the 6 dB attenuation of the whole sample. The upward-pointing triangles represent the amplifications and the downward-pointing triangles the attenuations depicted in Fig. 2. For modification set A, these are located on the horizontal axis at the time when the attenuation or amplification of the impulse response begins. For modification set B, the placement on the horizontal axis corresponds to the time when the attenuation or amplification ends. The triangles are slightly offset from their actual positions on the horizontal axis to avoid overlapping.

6. CONCLUSIONS

Many factors that affect auditory distance perception have previously been identified, one well-known factor being reverberation. The direct-to-reverberant energy ratio is often considered a measure of the effect that reverberation has on the perceived distance of a sound source. However, many details concerning the interaction between reverberation and auditory distance perception remain unclear. This study tries to shed some light on the matter, by looking at how modifications of the temporal envelope of binaural room impulse responses affect the perceived distance of virtual speech sources. The results suggest that the perceived distance can be more effectively controlled by modifying an early-to-late energy ratio, including approximately 50–100 ms of the impulse response in the early energy, than by directly modifying the traditional direct-to-reverberant energy ratio.

7. REFERENCES

[1] P. Zahorik, D. Brungart, and A. Bronkhorst, "Auditory distance perception in humans: A summary of past and present research," *Acta Acustica united with Acustica*, vol. 91, no. 3, pp. 409–420, 2005.

[2] J. M. Chowning, "The simulation of moving sound sources," *Journal of the Audio Engineering Society*, vol. 19, no. 1, pp. 2–6, 1971.

[3] P. Zahorik, "Direct-to-reverberant energy ratio sensitivity," *The Journal of the Acoustical Society of America*, vol. 112, no. 5, pp. 2110–2117, 2002.

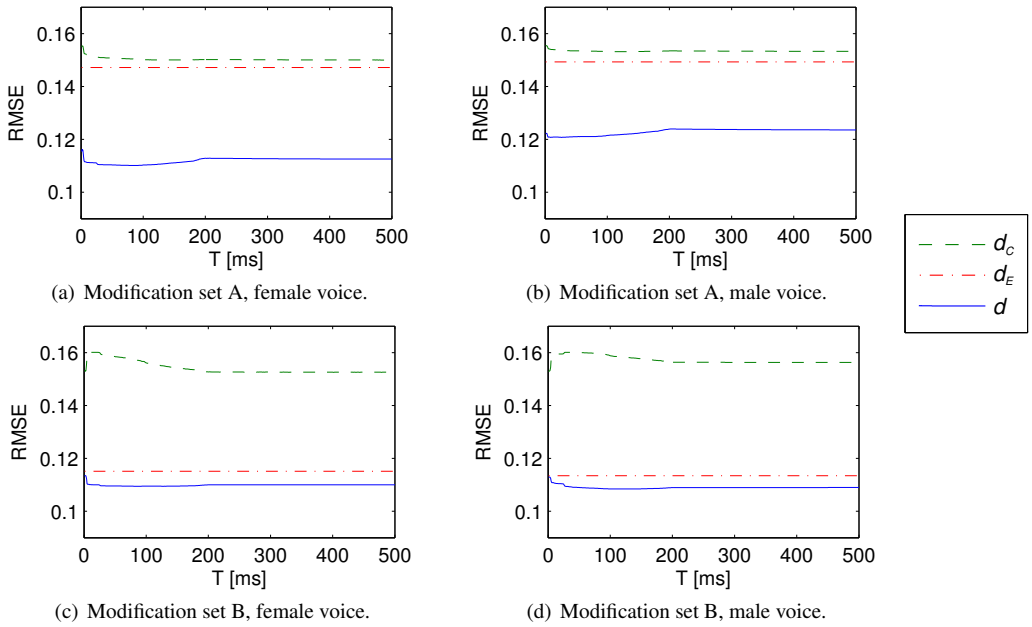
[4] E. Larsen, N. Iyer, C. R. Lansing, and A. S. Feng, "On the minimum audible difference in direct-to-reverberant energy ratio," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 450–461, 2008.

[5] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *The Journal of the Acoustical Society of America*, vol. 113, no. 6, pp. 3233–3244, 2003.

[6] A. Bronkhorst and T. Houtgast, "Auditory distance perception in rooms," *Nature*, vol. 397, no. 6719, pp. 517–520, 1999.

[7] D. R. Begault, "Preferred sound intensity increase for sensation of half distance," *Perceptual and Motor Skills*, vol. 72, no. 3, pp. 1019–1029, 1991.

[8] M. B. Gardner, "Distance estimation of 0° or apparent 0°-oriented speech signals in anechoic space," *The Journal of the Acoustical Society of America*, vol. 45, no. 1, pp. 47–53, 1969.



(a) Modification set A, female voice.

(b) Modification set A, male voice.

(c) Modification set B, female voice.

(d) Modification set B, male voice.

Figure 8: Root-mean-square error (RMSE) when fitting Eq. (6) (dashed line), Eq. (7) (dash-dot line), and Eq. (8) (solid line) to the distance estimates in space II, shown for different values of T.

[9] D. S. Brungart and K. R. Scott, "The effects of production and presentation level on the auditory distance perception of speech," *The Journal of the Acoustical Society of America*, vol. 110, no. 1, pp. 425–440, 2001.

[10] A. Bronkhorst, "Modeling auditory distance perception in rooms," in *Proceedings of EAA Forum Acusticum*, Sevilla, Spain, September 2002.

[11] D. S. Brungart, "Near-field virtual audio displays," *Presence: Teleoperators and Virtual Environments*, vol. 11, no. 1, pp. 93–106, 2002.

[12] D. S. Brungart and B. D. Simpson, "Auditory localization of nearby sources in a virtual audio display," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, USA, October 2001, pp. 107–110.

[13] E. R. Calcagno, E. L. Abregú, M. C. Eguía, and R. Vergara, "The role of vision in auditory distance perception," *Perception*, vol. 41, no. 2, pp. 175–192, 2012.

[14] D. H. Mershon and L. E. King, "Intensity and reverberation as factors in the auditory perception of egocentric distance," *Perception & Psychophysics*, vol. 18, no. 6, pp. 409–415, 1975.

[15] P. Zahorik, "Assessing auditory distance perception using virtual acoustics," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1832–1846, 2002.

[16] M. Jeub, M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proceedings of the 16th International Conference on Digital Signal Processing*, Santorini, Greece, July 2009.

[17] M. Jeub, M. Schäfer, H. Krüger, C. Nelke, C. Beaugeant, and P. Vary, "Do we need dereverberation for hand-held telephony?" in *Proceedings of the 20th International Congress on Acoustics*, Sydney, Australia, August 2010.

[18] D. Begault and E. Wenzel, "Headphone localization of speech," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 35, no. 2, pp. 361–376, 1993.

[19] P. Kabal, "TSP speech database," 2002. [Online]. Available: http://www-mmsp.ece.mcgill.ca/Documents/Downloads/TSPspeech/

[20] G. A. Soulodre and J. S. Bradley, "Subjective evaluation of new room acoustic measures," *Journal of the Acoustical Society of America*, vol. 98, no. 1, pp. 294–301, 1995.

[21] E. Skovenborg and S. H. Nielsen, "Evaluation of different loudness models with music and speech material," in *Audio Engineering Society Convention 117*, San Francisco, CA, USA, October 2004, paper no. 6234.

[22] V. P. Sivonen and W. Ellermeier, "Directional loudness in an anechoic sound field, head-related transfer functions, and binaural summation," *The Journal of the Acoustical Society of America*, vol. 119, no. 5, pp. 2965–2980, 2006.

[23] S. S. Stevens, "On the psychophysical law." *The Psychological Review*, vol. 64, no. 3, pp. 153–181, 1957.