Perceptual Evaluation of Model- and Signal-Based Predictors of the Mixing Time in Binaural Room Impulse Responses*

ALEXANDER LINDAU, *AES Student Member,* **LINDA KOSANKE, AND STEFAN WEINZIERL**
(alexander.lindau@tu-berlin.de) (l-kosanke@mailbox.tu-berlin.de) (stefan.weinzierl@tu-berlin.de)

Audio Communication Group, TU Berlin, Germany

The mixing time of room impulse responses denotes the moment when the diffuse reverberation tail begins. A diffuse ("mixed") sound field can physically be defined by (1) equidistribution of acoustical energy and (2) a uniform acoustical energy flux over the complete solid angle. Accordingly, the *perceptual* mixing time could be regarded as the moment when the diffuse tail cannot be distinguished from that of any other position or listener's orientation in the room. This, for instance, provides an opportunity for reducing the part of binaural room impulse responses that has to be updated dynamically in Virtual Acoustic Environments. Several authors proposed model- and signal-based estimators for the mixing time in rooms. Our study aims at an evaluation of all measures as predictors of a perceptual mixing time. Therefore, we collected binaural impulse response data sets with an adjustable head and torso simulator for a representative sample of rectangular shaped rooms. Altering the transition time into a homogeneous diffuse tail in real time in an adaptive, forced-choice listening test, we determined just audible perceptual mixing times. We evaluated the performance of all potential predictors by linear regression and finally obtained formulae to estimate the perceptual mixing time from measured impulse responses or physical properties of the room.

0 INTRODUCTION

Room impulse responses are usually considered to comprise three successive parts: direct sound, early reflections, and a tail of stochastic reverberation. The transition point between early reflections and the stochastic reverberation tail is often called mixing time (t_m) [1]. Due to increasing reflection density and diffuseness of the decaying sound field the perceptual sensitivity for the temporal and spectral structure of room impulse responses decreases during the decay process, and individual reflections become less and less distinguishable [2]-[4]. Moreover, auditory suppression effects, as, e.g., level-, direction-, and time-dependent simultaneous and nonsimultaneous masking or the precedence-effect further affect the audibility of reverberation fine structure. Further, Olive and Toole [5] pointed out the role of the audio content in discriminating room reflections.

Computational demands for Virtual Acoustic Environments (VAEs) will be reduced with the amount of early reflections to be rendered. A common method to achieve this is to replace the individual reverberation tail of binaural room impulse responses (BRIRs)—after an instant when *perceptual* discrimination is no longer possible—with an arbitrary and constant reverberation tail. Also for efficient loudspeaker array-based sound field synthesis it is relevant to know how many individual early reflections have to be rendered and when a stochastic tail is sufficient. In the following, this instant will be referred to as the perceptual mixing time $t_{mp.}$

Already in an early publication on dynamic auralization it was proposed to split the convolution process into a time variant convolution of the early impulse response (IR) parts and a static convolution with an arbitrary reverberation tail [2]. For pre-calculated BRIRs of a concert hall, in [2], this split point was set after 4000 samples corresponding to a transition point at 83 ms.

In [4], for a lecture hall (V = 420, RT = 1.0 s) and for less critical audio material, a split point of 80 ms was

*This paper is a revised and updated version of AES paper 8089, ''Perceptual Evaluation of Physical Predictors of the Mixing Time in Binaural Room Impulse Responses,'' recognized as the cowinner of the AES 128th Convention Student Technical Paper Award.

found sufficient for crossfading into an artificially designed reverberant tail.

For a small room ($V = 185\ m^3$, $RT = 0.72$ s), Meesawat and Hammershøi examined t_{mp} for different combinations of early reflections and diffuse tails [3]. Manipulating the crossfade instances authors determined perceptual mixing times for (a) interchanged tails of the two ears, (b) tails from different receiver positions while keeping the same relative source position, (c) tails from the same receiver position but different horizontal source angles, and finally (d) tails from different receiver *and* source positions. Stimuli were convolved with accordingly manipulated binaural impulse responses, resulting in static auralization for presentation. For a listening test, the method of constant stimuli was used. Results lead to the conclusion that—for this room—t_{mp} was about 40 ms and independent from all position changes tested.

Since we expected higher mixing times for larger rooms, in a past study [6], we assessed t_{mp} for a large auditorium ($V = 8500\ m^3$, $RT = 2$ s), also using static auralization but an adaptive listening test design. The perceptual mixing time was indeed higher (up to 140 ms). In addition to findings in [3], we found no effect of taking a tail from the same receiver position but for different head orientations. However, the effect of taking the tail from a different source position *and* a different head orientation (case not tested by Meesawat and Hammershøi, [3]) led to considerably increased perceptual mixing times; potential reasons will be discussed in Section 0.2. Additionally, it turned out that listeners were most sensitive when a specific drum set sample with strong transients was used.

The aim of our present study was to find the perceptual mixing times for approximately shoebox shaped rooms of differing volume and average absorption while utilizing state of the art dynamic auralization and an adaptive, forced-choice listening test design. Subsequently, we examined several predictors of the physical mixing time for their ability to predict the perceptual mixing time.

0.1 The Concept of Physical Mixing Time

Diffusion and *mixing* are often used synonymously for the characteristic of sound fields in real enclosures to become more and more stochastic over time.

The transition from early reflections into a stochastic reverberation tail is a gradual one. Every time a sound wave hits a wall, it is reflected. Depending on the surface properties, this reflection can be specular, partly, or fully diffuse. In an ideally diffuse reflecting room, the sound energy continuously spreads over the whole volume in time. Finally, the ideal diffuse sound field is characterized by a uniform angular distribution of sound energy flux and a constant acoustical energy density over the whole space [8]. The process of mixing, on the other hand, is usually illustrated in terms of particle trajectories, when, over time, position and direction of two initially adjacent rays become statistically independent. A requirement for a room to become mixed is ergodicity, which means that—at some point in time—the statistical behavior at all points in the space equals that at one point over time (time average equals ensemble average, [1], [9]), i.e., the sound field has completely "lost any memory" of its initial state.

The duration of the diffusion process, i.e., the physical mixing time, increases with room size as—due to larger free path lengths— the time intervals between individual reflections are increased. This effect is further pronounced if the room is lacking any diffusing obstacles [7].

0.2 Real-World Limitations of Complete Mixing

Ergodicity was shown to be dependent on the shape of the enclosure and the surface reflection properties [10]. Examples for non-ergodic rooms are perfectly rectangular non-diffusing rooms (particle directions remain deterministic) or non-diffusing spherical rooms (due to focusing not all positions will be reached by a particle). The process of mixing within the decaying room impulse response may further be disturbed in rooms with non-uniform distribution of large areas with highly varying absorption coefficients (for instance, when large windowpanes are combined with highly absorbing audience seats). As shown by Pollack [1], absorbing rooms can never be perfectly diffuse, because there always remains a net energy flow in the direction of the losses (i.e., toward the absorbing walls). Also coupled rooms, highly damped rooms, and very small rooms may lack mixing in their decay.

Inherently, the whole concept of mixing is further confined to a frequency range where the theory of geometrical and statistical acoustics applies. In real rooms, these assumptions are violated by modal behavior in the low-frequency range. Another problem might arise from proximity to room boundaries (sidewalls, floors). In this case reflections may form distinct comb filters whose spectra depend on the exact receiving position violating the assumption of positional independence of the diffuse sound field. In summary, it can be stated that perfect mixing (or total diffusion) is an idealization never fully encountered in real rooms.

0.3 Physical and Perceptual Mixing Time

With reference to the physical definition we propose to say that a room is perceptually mixed, if the stochastic decay process at one position in the enclosure cannot be distinguished from that of any other position and/or listener orientation. Due to auditory and cognitive suppression effects as well as properties of the audio content mentioned above, the perceptual mixing time can expected to be equal or smaller than the physical mixing time, no matter how the latter is determined. An operationalization of mixing time still has to be defined and should take into account the intended application. Below, we will introduce an experimental concept aiming at applications in binaural technology.

1 METHODS

Section 1.1 recapitulates common model-based estimators of the mixing time. Section 1.2 gives an overview of

four recently proposed signal-based parameters. Section 1.3 explains the motivation for selecting the rooms for the listening tests, whereas Section 1.4 describes the measurement of binaural room impulse responses for the listening test. Section 1.5 gives details about the actual calculation of these parameters and the treatment of practical issues we encountered. Section 1.6 explains the listening test, and finally, Section 1.7 explains the listener selection procedure.

1.1 Model-Based Estimators of Mixing Time

Several estimators of the perceptual mixing time t_{mp} have been suggested in literature. Ad hoc values as for instance 50 ms [11], or 80 ms ([2], [12]) have been proposed regardless of further room properties. Other authors suggest time ranges of 100–150 ms [13], 150–200 ms [14], or 50–200 ms [15], to take into account different room properties.

Some theoretically motivated estimators of t_{mp} explicitly refer to properties of the auditory system such as time resolution or being "free from flutter" and assume reflection densities from 250 s^{-1} [17], 400 s^{-1} ([14], [18]), 1000 s^{-1} [19], 2000 s^{-1} [20], 4000 s^{-1} [21] up to 10.000 s^{-1} [22] to be sufficient to render stochastic reverberation tails. Setting the reflection density *dN/dt* as derived from the mirror source model of the rectangular room

$$\frac{dN}{dt} = \frac{4\pi \cdot c_0^3 \cdot t^2}{V} \tag{1}$$

(c_0: sound velocity in m/s, V: room volume in m^3) to 400 s^{-1}, and solving for t, the following popular estimation of the mixing time was proposed in [16]:

$$t_{mp} \approx \sqrt{V}, \text{ with } t_{\mathrm{mp}} \text{ in ms.} \tag{2}$$

Rubak and Johansen [21] introduced a different view, relating the instant of perceptual mixing to the concept of the mean free path length l_{m}:

$$l_m = 4\frac{V}{S}, \tag{3}$$

where S is the total surface area of the enclosure in m^2. The rationale of this approach is that the sound field is assumed to be virtually diffuse, if every sound particle has on average undergone at least some (e.g., [21]: four) reflections. Thus their estimation of t_{mp} (in ms) reads:

$$t_{mp} \approx 4l_m \frac{10^3}{c_0} = 4 \cdot \left(\frac{4V}{S}\right) \cdot \frac{10^3}{c_0} \approx 47 \cdot \frac{V}{S}. \tag{4}$$

Recently, Hidaka et al. [23] proposed a linear regression formula that fits results from a larger study on physical mixing times determined empirically from impulse responses, including 59 concert halls of different shape and size. The formula predicts the mixing time t_{m500Hz} for the 500 Hz octave band (in ms) from the room's reverberation time

$$t_{m500Hz} = 80 \cdot RT_{500Hz} \tag{5}$$

Thus, all suggested estimators depend on only three room-specific quantities: volume, surface area, and reverberation time. They can therefore further be generalized to

$$t_{mp1} = k_{refl} \cdot \sqrt{V}, \tag{6}$$

for the reflection density relation (2). The general mean free path length relation similarly reads

$$t_{mp2} = k_{path} \cdot \frac{V}{S}, \tag{7}$$

and the general estimation from the reverberation time can be rewritten as

$$t_{mp3} = k_{reverb} \cdot RT_i. \tag{8}$$

Thus, three basic model-based relations of the mixing time remain to be subjected to a perceptual evaluation.

1.2. Signal-Based Predictors of Physical Mixing Time

Recently, several algorithms were proposed for calculating the physical mixing time from empirical room impulse responses. We included four of these approaches into our evaluation.

Abel and Huang (2006)

Abel and Huang [25] proposed an approach based on the assumption that the sound pressure amplitudes in a reverberant sound field assume a Gaussian distribution. For determining the mixing time, a so-called "echo density profile" is calculated. With a short sliding rectangular window of 500–2000 samples, the empirical standard deviation of the sound pressure amplitudes is calculated for each sample index. In order to determine how well the empirical amplitude distribution approximates a Gaussian behavior, the proportion of samples outside the empirical standard deviation is determined and compared to the proportion expected for a Gaussian distribution. With increasing time and diffusion, this echo density profile should increase until it finally—at the instant of complete diffusion— reaches unity. With larger window sizes, the overall shape of the echo density profile stays similar whereas smoothing of the fine structure can be observed. We chose a rectangular window of 2^{10} samples (23 ms), as suggested by the authors while referring to the auditory temporal resolution. The mixing time can be defined as the instant where the echo density profile becomes unity for the first time (criterion I). In order to account for minor fluctuations Abel and Huang modified this criterion to account for the instant when the reflection density is within *1*-σ_{late} (σ_{late} being the standard deviation of the late echo density, criterion II*)*. We evaluated both stopping criteria, while calculating σ_{late} from the last 20% of the impulse responses before reaching the noise floor.

Stewart and Sandler (2007)

Following an idea proposed in [25], Stewart and Sandler [26] suggested measuring the kurtosis of the sound pressure amplitudes and comparing this value to that expected for a Gaussian distribution. As a second order cumulant, the kurtosis γ_4 is a measure of the "non-Gaussianess" contained in a signal. In the normalized form, $\gamma_{4\mathrm{n}}$ is given as:

$$\gamma_{4n} = \frac{E\{x - \mu\}^4}{\sigma^4} - 3. \tag{9}$$

where E is the expectation operator, μ is the mean, and σ is the standard deviation of the process. For increasingly Gaussian-like processes, the normalized kurtosis must approach zero. We calculated this instant with identical settings as for the echo density profile. Although not clearly stated in [26], we concluded from the authors' discussion, that the instant when the normalized kurtosis γ_{4n} reached zero for the first time should assumed to be the mixing time.

Hidaka et al. (2007)

Hidaka et al. [23] proposed a frequency-domain approach for the estimation of the instant when a room impulse response has become diffuse. Therefore, the time-frequency energy distribution of the impulse response *p(t)* is calculated according to

$$E(t, \omega) = \left| \int_{t}^{\infty} p(\tau) e^{j\omega\tau} d\tau \right|^2 . \quad (10)$$

When averaging over a frequency range $\Delta\omega$, (10) can be shown to be identical to the Schroeder integration [24]. The energy distribution *E(t,ω)* is calculated for impulse responses beginning with the direct sound; initial delays are removed in advance. With increasing time *t, E(t,ω)* will progressively contain fewer early reflections and more stochastic reverberation. In a second step, Pearson's product-moment correlation *r(t)* is calculated as a continuous function of time for $E(0{:}\infty, \Delta\omega)$ and $E(t{:}\infty, \Delta\omega)$ in arbitrary frequency bands. This will describe the similarity between (a) the energy decay process including the initial state and (b) the energy decay process with beginning from any time t afterward in one particular frequency band. Hidaka et al. define the "transition time" into stochastic reverberation as the instant when $r(t) \leq e^{-1} = 0.368$. Thus, we calculated *E(t,ω)* and *r(t)* for octave bands from 125 Hz to 16 kHz and detected the mixing time at the moment when $r(t) \leq 0.368$ for the first time. For ease of computation we limited the temporal resolution to 100 samples ($\Delta t = 2.3$ ms).

Defrance et al. (2009)

Recently, Defrance et al. [27] suggested a new procedure for estimating the physical mixing time from room impulse responses. Their method is based on the assumption that, over time, the reflection density at an observing point in an enclosure becomes so large, that singular reflections begin to overlap and cannot be distinguished anymore. The authors propose a technique ("Matching Pursuit") somewhat similar to wavelet decomposition to decompose room impulse responses into singular reflections (called "arrivals"). As a result they obtain a function of the cumulative number of arrivals, which—as can be derived from time integration of (1)—should show a cubic increase. Following the authors' argumentation the decomposition process should more and more fail to distinguish superimposed reflections resulting in a slope changing from cubic to linear. The instant of the changing slopes would then equal the physical mixing time.

Considering all reflections to be more or less copies of the direct sound impulse only the direct sound itself is used as wavelet in the "Matching Pursuit." Decomposition is conducted by correlating the impulse response with the direct sound while shifting the latter along the impulse response to all possible instances in time. At the instance of maximum correlation, the direct sound is subtracted from the IR weighted by the corresponding correlation value. The decomposition process is repeated until the energy ratio SRR (signal residual ratio) of the reconstructed signal (reconstructed from the direct sound and the time vector of correlation coefficients) and remaining impulse response signal (the residuum) rises above a certain value. To avoid a decomposition wrongly favoring the early parts of the IR, its energy decay has to be compensated before running the decomposition. Finally, the mixing time is determined by applying a reflection distance criterion on the reconstructed impulse response. Therefore, Defrance et al. argued that the mixing time would be equivalent to the moment were the first two reflections are spaced equal or less than the so-called "equivalent duration" of the direct sound.

Using the software provided by the authors, we were able to calculate the Matching Pursuit decomposition using their original Matlab® code. Additionally, we implemented the energy decay compensation and the calculation of the equivalent duration of the direct sound. According to recommendations in [27] we used a SRR of 5 dB as stopping criterion for all decompositions.

1.3. Room Selection

The main purpose of this study was to find reliable predictors for the perceptual mixing times for a broad range of acoustical environments. The physical mixing times derived in [23] for a large selection of concert halls were at maximum for shoebox shaped rooms. From the theory of mixing (see Section 0.2), their regular shape and their long unobstructed path lengths suggests them to be most critical in terms of late mixing times. Therefore, we confined this study to largely rectangular rooms. Coupled enclosures were avoided. Wall surface materials varied from only little diffusing concrete, glass or gypsum to considerably structured wood panels. Floors were made from linoleum, parquet or granite. The smaller (lecture) rooms were equipped with chairs and tables, whereas all the larger rooms included extended audience seating areas. For the sake of simplicity, we calculated surface area from the three main dimensions of the considered ideal shoebox room, neglecting additional surfaces of galleries or furniture.

We selected nine rooms, aiming at a systematic variation of both volume and average absorption coefficient (α_{avg}), each in three steps (see Table 1). This so-called complete variation would permit an independent statistical assessment of both influences by means of two-way ANOVA.

Due to physical interrelation, it is difficult to vary room volume independently from *absolute* amount of reverberation, i.e., from the reverberation time. However, while varying the average absorption coefficient, we could at least assess the influence of the *relative* amount

Table 1. Volume, average absorption coefficient, and reverberation time of the nine selected rooms.

	large α (RT)	med. α (RT)	small α (RT)	avg. Vol.
small V	**room 1** 216 m^3 α: 0.36 (0.39 s)	**room 2** 224 m^3 α: 0.26 (0.62 s)	**room 3** 182 m^3 α: 0.17 (0.79 s)	**207 m^3**
medium V	**room 4** 3300 m^3 α: 0.28 (1.15 s)	**room 5** 5179 m^3 α: 0.23 (1.67 s)	**room 6** 3647 m^3 α: 0.2 (1.83 s)	**4042 m^3**
large V	**room 7** 8298 m^3 α: 0.33 (1.52 s)	**room 8** 8500 m^3 α: 0.23 (2.08 s)	**room 9** 7417 m^3 α: 0.23 (2.36 s)	**8072 m^3**
avg. α (RT)	**0.32** (1 s)	**0.24** (1.45 s)	**0.2** (1.66 s)	

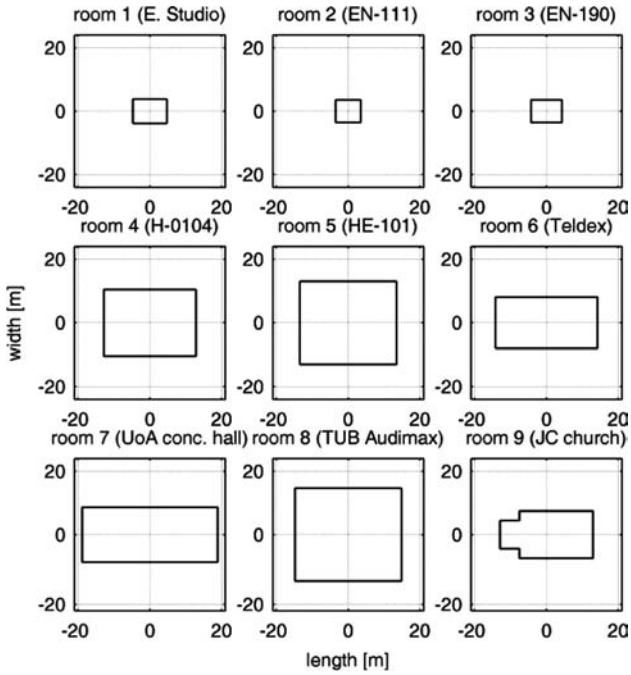

Fig. 1. Floor plans of the nine rooms (true scale).

of reverberation independently of volume. Step sizes of reverberation time were additionally chosen to exceed at least a just noticeable difference of 10%. The most important parameters of the selected rooms are listed in Table 1. Additionally, Fig. 1 shows true to scale floor plans of the rooms.

The three small rooms were, in order of increasing reverberation time, the Electronic Studio of the TU Berlin (room 1), and two small lecture rooms, EN-111 and EN-190 (room 2 and 3). The medium size rooms were the TUB's lecture halls H-104 (room 4), and HE-101 (room 5), and the large recording room of the Teldex Studio Berlin (room 6). The three large venues comprised the concert hall of the University of Arts in Berlin (room 7), the auditorium maximum of the TUB (room 8), and the Jesus-Christus Church in Berlin-Dahlem (room 9). Rooms 4, 6, 7, 8, and 9 are regularly used as musical performance spaces.

1.4. Binaural Measurements

In order to provide high quality dynamic binaural simulations of all environments, we measured binaural room impulse responses in all nine rooms using the automatic head and torso simulator FABIAN [6].

As sound source, a 3-way dodecahedron loudspeaker, providing high signal to noise ratio and optimal omnidirectional directivity, was placed in the middle of the stage, which was typically located at one of the narrow ends of the rooms. To obtain a wide frequency range for the BRIRs, the loudspeaker was equalized to a linear frequency response within ± 3 dB from 40 Hz to 17 kHz. In a first step, monaural room impulse responses were measured at three different positions in the diffuse field using an omnidirectional microphone. Second, the three major room dimensions length, width, height were measured for calculating the volume. Then, the reverberation times displayed in Table 1 were calculated in-situ as an average over the octave bands from 125 Hz to 4 kHz and all three measurement positions. Now, the critical distance could be derived and FABIAN was seated on a place on the room's longitudinal symmetry axis directly facing the loudspeaker at twice the critical distance, where, according to Kuttruff [28], a random sound field can be expected. BRIRs were collected for horizontal head orientations within ±80° in angular steps of 1°.

1.5. Practical Considerations when Calculating Signal-Based Mixing Time Parameters

We calculated the signal-based mixing time parameters from the dummy head's left and right ears' impulse response of the neutral head orientation (i.e., when facing the sound source) and from the three measurements collected with the omnidirectional microphone resulting in five signal-based mixing time estimates for each room. We considered all four approaches introduced in Section 1.2. The mixing time according to Abel and Huang [25] was calculated using both mentioned stopping criteria I and II. The estimator proposed by Hidaka et al. [23] was calculated individually for the eight octave bands between 125 Hz–16 kHz.

A major issue observed with signal-based parameters was the variability of t_m values with measurement position. All mixing times derived from the four approaches are depicted in Fig. 2 (Abel and Huang only for criterion I, Hidaka et al. only for 500 Hz octave band).

As can be seen, values vary by factors up to two or three and even more in, e.g., the case of Defrance et al.

As discussed already in Section 0.2 such positional variances might indicate imperfect mixing caused by close room boundaries, changing low frequency modal patterns, or residual coupled volumes. Positional variability of measures has partly been subject of discussion in the original publications and is not uncommon for certain room acoustical parameters (see e.g., [29]). Since for BRIRs there are always two impulse responses available, we also tried to reduce this variability by subjecting the mean of the mixing

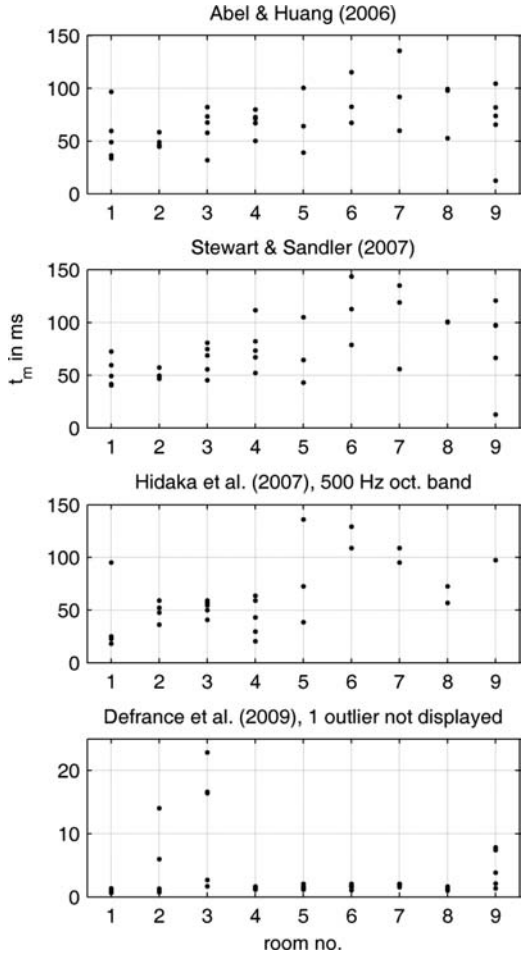

Fig. 2. Positional variability obtained with automatic t_{m} detection using signal-based parameters (five values per room: left and right ears BRIR, three IRs from omnidirectional microphone measurements).

time values as estimated from both channels of a BRIR to later statistical analysis.

Moreover, in case of the results from the Matching Pursuit decomposition [27], most values determined automatically were implausibly low (often around 2–5 ms, see Fig. 2, bottom). This behavior was described already by Defrance et al. [27], when discussing the dependency of the estimates and their spread on the chosen signal residual ratio (SRR) in the Matching Pursuit decomposition (see also Section 1.2).

Apart from purely physical explanations for the observed positional variability, from examination of the plots of echo density profiles, normalized kurtosis or cumulated arrival function (see Fig. 3) we suspected the different criteria for determination of the mixing time to be another reason for the positional instability of t_m measures. As can be seen from Fig. 3 (upper and middle plot), echo density and normalized kurtosis did not always approach their target values (1 or 0, resp.) in a continuous manner, but jumped occasionally.

Thus, measuring the time until the target value is reached for the first time might not always be a reasonable criterion. Besides, from parameters' profile plots of Abel

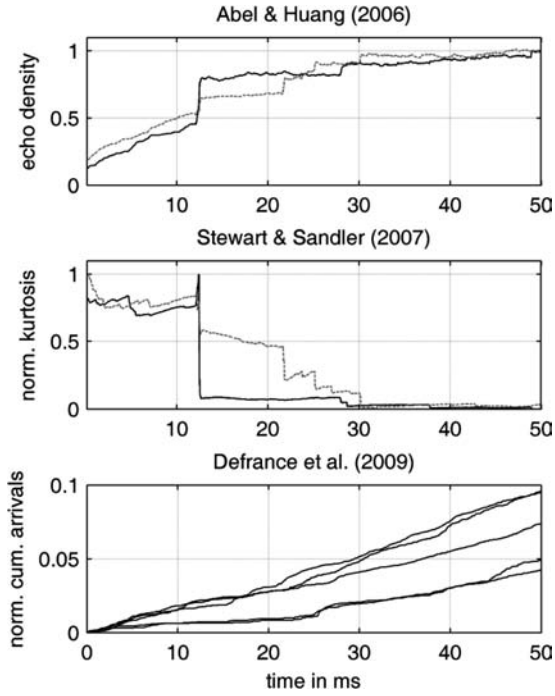

Fig. 3. Positional variability of signal-based parameters' profiles (plots 1 and 2: from both BRIR channels in room 2; plot 3: from all five impulse responses in room 5).

and Huang's, Stewart and Sandler's, and Defrance et al.'s method the mixing time could also be determined visually. In reading off mixing times from the location of the last noticeable jump toward the target value in the profile plots of echo density or normalized kurtosis, respectively, we hoped to get more stable results. Regarding Defrance et al.'s method, reading off the point where the slope of the cumulated arrival function changes from cubic to linear was often difficult (see Fig. 3, bottom).

Despite these problems, for the three approaches we additionally determined mixing time values by visual inspection of the corresponding curves and subjected them to our later statistical analysis.

1.6. Listening Test

Perceptual mixing times were determined using an adaptive 3-AFC (three-alternative forced-choice) listening test procedure. Subjects were confronted with three stimuli in random order. Two of them were the reference simulation, where the complete BRIRs were updated in real time according to head movements. One of them was the manipulated simulation, where only the early part of the BRIRs was dynamically updated, whereas the late reverberation tail—taken from the BRIR corresponding to frontal head orientation—was concatenated with a linear crossfade within a window size corresponding to the early block size of the fast convolution engine. This block size also corresponded to the step width the mixing time could be altered with (see Fig. 4).

Thus, the concatenation point of the updated early and the static late BRIR could be changed in increments of 5.8 ms (small rooms, nos. 1 to 3), or 11.6 ms (medium and large

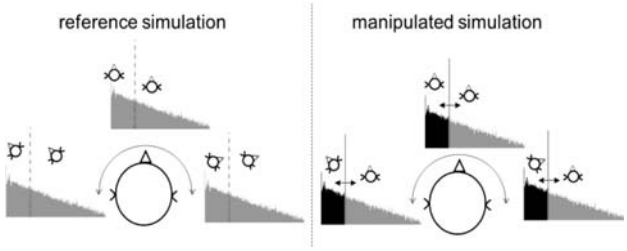

Fig. 4. The two stimulus conditions presented in the listening test. Reference sound field/left: The complete BRIR is continuously updated according to the current head orientation. Manipulated sound field/right: Only the early BRIR part is continuously updated; the late reverberation always corresponds to frontal head orientation. The concatenation point between early and late part of the BRIRs was adaptively altered in the listening test.

rooms, nos. 4 to 9). Whenever the subjects could correctly identify the manipulated simulation, the duration of the early dynamic part of the BRIR was increased; otherwise it was reduced, forcing the transition time to converge to the just noticeable point.

Following the definition proposed in Section 0.3, if the rooms were totally mixed everywhere and in the complete frequency range, BRIRs from every position in the rooms could have delivered the static reverberant tail. When assessing reverberant tails from different measurement positions in pretests, however, low frequency differences between original and manipulated tails were clearly audible, leading to high perceptual mixing times. As discussed in Section 0.2, different effects may have disturbed mixing of the sound field, such as comb filters caused by near room boundaries or a position-dependent low frequency response of the rooms due to modal patterns below the Schroeder frequency. Hence, the BRIRs of the neutral head orientation (i.e., from the same location as the dynamic BRIR datasets used for auralization) were used in order to avoid positional dependencies still observable in the late stochastic tail.

Although the noise floor was below –80 dB relative to the direct sound for all measurements, we limited the length of the BRIRs to about three-quarters of the duration of the decay to noise floor, because a slightly different background noise level and spectral coloration for different BRIRs was audible when comparing reference and manipulated stimuli and would thus have biased the detection task. Hence, BRIRs had a length between 14.000 (i.e., 0.375 s for room 1) and 100.000 (i.e., 2.26 s for room 9) samples, maintaining approximately 60 dB decay for binaural simulations of all rooms.

Loudness differences between simulated rooms were minimized through normalization of BRIR datasets. Electrostatic headphones (STAX SR-2050II) were frequency compensated using fast frequency deconvolution with high pass regularization from measurements on our dummy head FABIAN [30], [31]. Subjects were allowed to adjust sound pressure during training to a convenient level. This level was then kept constant throughout the listening test.

The listening test was conducted using the WhisPER toolbox [32]. As adaptation method for the threshold value (here: the just audible transition time), a Bayesian approach that closely matches the ZEST procedure [33] was chosen due to its unbiased and reliable results. The a-priori probability density function was a Gaussian distribution with its mean in the middle of the stimulus range; the standard deviation was adapted in pretests.

As stimulus, the critical drum set sample from [6] was used again (length: 2.5 s without reverberation tail). The three stimuli where played back successively in each trial, without the possibility to repeat a trial. Subjects had to assess all nine rooms in an individually randomized order. A run was stopped after 20 trials, resulting in a test duration of about 60 minutes per person.

1.7. Subjects

During pretests, the listening test turned out to be a difficult task for some subjects. Consequently, we introduced a criterion to select "expert listeners" as those who were able to detect the right perceptual cue in order to perform the difference detection task successfully. Therefore, we regarded those subjects as experts, who were able to achieve, in all of the nine tested rooms, thresholds that were larger than the earliest possible concatenation instant (i.e., when concatenating only the dynamically updated direct sound with a static diffuse tail).

Finally, results of 10 expert listeners (2 female, 8 male) with an average age of 27.8 were taken into account for further statistical analysis. Most of the subjects had a musical education background and all had participated in listening tests before. During training subjects were instructed to rotate their head widely for maximizing the difference between original and manipulated reverberation tails. To increase statistical power we used a repeated measures design (every subject assessed every stimulus condition).

2 LISTENING TEST RESULTS

For each room the just detectable perceptual mixing times t_{mp} were calculated as the moment corresponding to the middle of the crossfade window between early and late BRIR at the cross fade instant the adaptive algorithm had converged to after 20 trials. Fig. 5 shows the average perceptual mixing times t_{mp50} and confidence intervals ordered according to the two tested conditions volume and average absorption coefficient. As expected, t_{mp50}-values

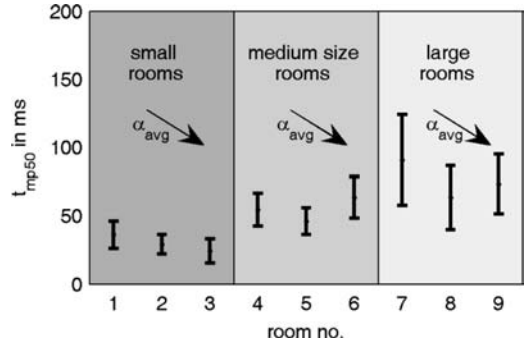

Fig. 5. Average perceptual mixing times t_{mp50} per room with 95% CIs.

were found to increase with room volume. As indicated by the growing confidence intervals of rooms 7–9 the variation among subjects increased, too.

The ANOVA for repeated measures proved the volume effect to be significant at $p = 0.001$. Trend analysis confirmed a significant positive linear relation. An effect of the average absorption coefficient (i.e., the relative reverberance independent of volume) could not be found. This is in accordance with theory, as the amount of reverberation is in principle not related to diffusion or mixing. However, our sample size allowed only testing of intermediate effects ($f_{posthoc} = 0.307$).

3 REGRESSION ANALYSIS

In Section 3.1, results of a regression analysis conducted to test the power of the three most important model-based relations (6)–(8) to predict t_{mp}, are presented. In Section 3.2, regression results for the signal-based parameters are discussed. In both cases, regression analysis was conducted for the average perceptual mixing time (t_{mp50}). Additionally, regression analysis was conducted while regressing on the 95%-point of the assumed normal distribution of the listening test results (t_{mp95}). While the t_{mp95}-regression formulae are intended to guarantee a perceptively perfect, close-to-authentic simulation, the t_{mp50}-predictions will guarantee a transparent simulation for at least half of the expert listeners.

Linear regressions were calculated as the least squares fit of the empirical t_{mp} values, derived in the listening test, to the t_m values as predicted by the above introduced model- and signal-based predictors. Thus, models of the form

$$t_{mp} = b_1 t_m + b_2 \tag{11}$$

were derived.

Although the intercept term b_2 cannot be easily interpreted in physical terms (a zero physical mixing time should predict a zero perceptual mixing time), a higher explained variance was obtained by allowing an intercept term.

All regression results were evaluated by means of the explained variance R^2, and the significance of regression, i.e., the possibility of rejecting the null hypothesis of a zero slope value b_1 at $p = 0.05$.

The number of nine rooms was rather low for linear regression analyses, a fact that is reflected by the confidence intervals displayed with the models. This shortcoming is, however, counterbalanced by the systematic and wide variation applied in selecting the rooms and by the selection of expert listeners, yielding a reliable measurement of perceptual mixing times with low variance among subjects.

3.1. Regression Results for Model-Based Predictors of Physical Mixing Time

As model-based predictors of the perceptual mixing time (a) the square root of volume, (b) the mean free path length, and (c) the reverberation time were subjected to regression analysis. Additionally, we tested the volume V, the surface area S (calculated from the three major room dimensions) and the average absorption coefficient α_{mean}.

Stepwise multiple and univariate linear regression analyses were conducted. Depending on the selection of variables, models containing one or three predictors resulted. The latter could be rejected, as the additional linear coefficients were insignificant, exhibited collinearity problems (high intercorrelation), and confidence intervals spanning to zero.

Thus, t_{mp50} was best predicted by the ratio V/S, the kernel of the mean free path length formula (7). In this case, the explained variance R^2 reached 81.5%. Regression on $\sqrt{V}$ (i.e., the reflection density relation) reached 78.6%, whereas volume alone achieved an R^2 of 77.4%. The reverberation time turns out to be unsuitable as predictor of the perceptual mixing time, since the explained variance of 53.4% can be completely attributed to confounded volume variation, while the average absorption coefficient α_{mean} shows nearly no linear relation to t_{mp50} ($R^2 = 0.8\%$). All regressions were significant, except the one derived for α_{mean}.

Fig. 6 shows t_{mp50} values and linear regression models including 95% confidence intervals of both data and models. The regression formula for the best predictor of t_{mp50} (in ms) was:

$$t_{mp50} = 20 \cdot V / S + 12 \tag{12}$$

Thus, when comparing (4) and (12) and neglecting the constant term of the regression model, one derives that after approximately two reflections sound fields were experienced as being diffuse. Additionally—and while also neglecting the constant model term— from the second best predictor found, a just audible reflection density can be estimated by substituting t in (1) with the first addend of the regression formula

$$t_{mp50} = 0.58 \cdot \sqrt{V} + 21.2. \tag{13}$$

Thus, with $c_0 = 343$ m/s, the just audible reflection density can be estimated as $dN/dt = 171\ s^{-1}$. These values are considerably lower than those traditionally suggested in the literature (see Section 1.1).

However, it must be emphasized, that these are not measured physical quantities but are inferred from the model-based relations (6) and (7). Moreover, the inferred just audible quantities might be true only in the case of large rooms where the neglected constant term of the linear models becomes more and more irrelevant.

When regressing on the stricter t_{mp95} values all models were significant, too, except for the one derived from the average absorption coefficient α_{mean}. The perceptual mixing time t_{mp95} (in ms) was best predicted by volume ($R^2 = 78.7\%$):

$$t_{mp95} = 0.0117 \cdot V + 50.1 \tag{14}$$

Results for further parameters are displayed in Table 2.

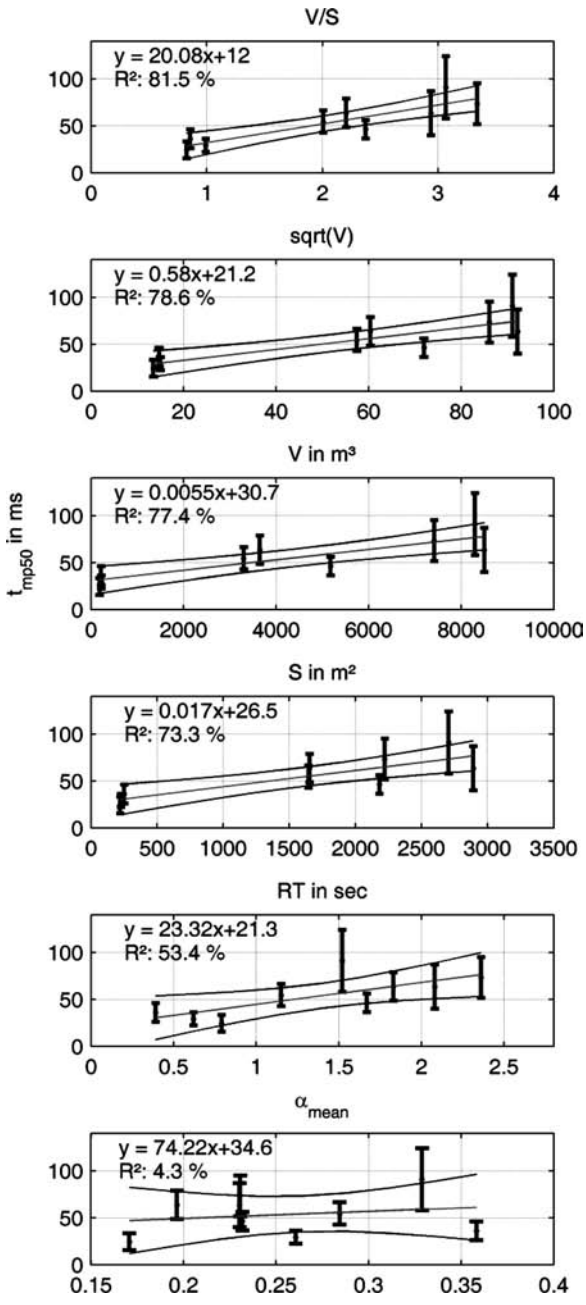

Fig. 6. Average perceptual mixing times t_{mp50} in ms (incl. 95% CIs) plotted over different model-based predictors, and resulting linear regression models (incl. hyperbolic 95% CI curves).

3.2. Regression Results for Signal-Based Predictors of Physical Mixing Time

Both the mixing time values calculated from the left and right ears' BRIR and their average, determined either (a) visually, or (b) using the described deterministic detection criteria (see Section 1.5) were subjected to linear regression analyses. Again, regression analysis was conducted for the average (t_{mp50}) and stricter perceptual mixing times (t_{mp95}).

For the algorithm of Defrance et al. [27], most of the estimated mixing time values were implausibly low (see Fig. 2), especially when assuming the signal-based approaches to be directly estimating the physical mixing time. A comparative visual inspection of the cumulative arrival functions suggests that the equivalent pulse duration criterion does not always lead to a correct detection of the inflection point of the cumulative arrival function. Therefore, and although in principal we consider this method as an attractive approach, we did not subject its results to regression analysis.

Table 2. Ranking of model- and signal-based mixing time estimators in predicting perceptual mixing times t_{mp50} and t_{mp95}.

	Model-based predictors		Signal-based predictors	
#	t_{mp50}	t_{mp95}	t_{mp50}	t_{mp95}
1	V/S R^2 81.5 %	V R^2 78.7 %	Abel I R^2 74.7 %	Abel I R^2 83.7 %
2	$\sqrt{V}$ R^2 78.6 %	V/S R^2 75.7 %	Hidaka 1k R^2 57.3 %	Abel II R^2 66.7 %
3	V R^2 77.4 %	$\sqrt{V}$ R^2 73.4 %	Hidaka 500 R^2 56.5 %	Hidaka 1k R^2 55 %
4	S R^2 73.3 %	S R^2 69.5 %	Abel II R^2 50.7 %	Hidaka 500 R^2 49.2 %
5	RT R^2 53.4 %	RT R^2 46.5 %	Stewart R^2 37.6 %	Stewart R^2 40.3 %
6	α_{mean} R^2 4.3 %	α_{mean} R^2 4.8 %		

Although some of the regression models derived from values of a single ear's BRIR reached higher values of explained variance, this happened randomly for the left or the right ear. A systematic relation could not be found, thus all further results are solely based on the average mixing time calculated from both channels of the BRIRs.

All regression models were significant, except the one derived from Stewart's and Sandler's [26] kurtosis-based method. The echo density approach of Abel and Huang [25] (criterion I) achieved a R^2 of 74.7% (see Fig. 7, plot 1). The obtained regression formula was

$$t_{mp50} = 0.8 \cdot t_{mix\_Abel\_I} - 8 \tag{15}$$

Therefore, we can recommend this estimator for signal-based determination of t_{mp50}. Regression models of the other approaches are depicted in Fig. 7, where results are presented in descending order of performance. The correlation approach from Hidaka et al. [23] reaches minor prediction performance but at least R^2 values of 56.5% to 57.3% for the mid frequency octave bands (500 Hz and 1 kHz).

Visually reading off mixing time values from the profile plots of (a) reflection density, (b) normalized kurtosis, or (c) cumulative arrivals resulted in regression models with considerably less explained variance. Moreover, as this procedure is very time consuming it cannot be recommended.

Prediction results for t_{mp95} are also shown in Table 2 ordered for performance. Again, all models were significant, despite that one derived from the kurtosis approach [26]. The approach of Abel and Huang again showed a superior performance, with an explained variance of 83.7%.

The regression formula reads

$$t_{mp95} = 1.8 \cdot t_{mix\_Abel\_I} - 38. \tag{16}$$

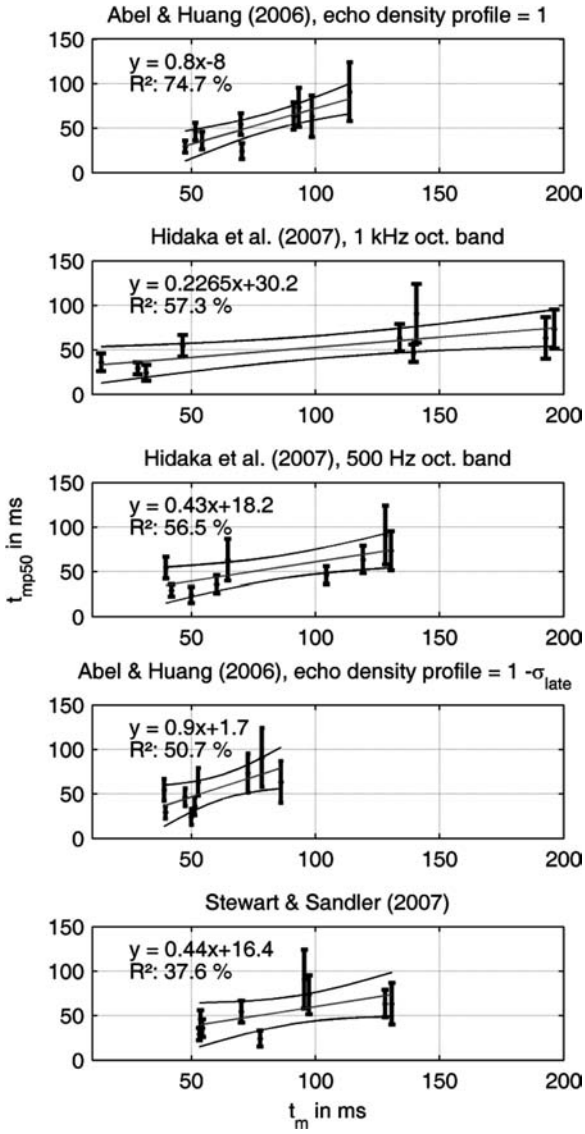

Fig. 7. Average perceptual mixing times t_{mp50} in ms (incl. 95% CIs) plotted over signal-based t_m-predictors (mean values from both channels of a BRIR), and resulting linear regression models (incl. hyperbolic 95% CI curves).

Further measures performed less well, though the echo density with criterion II [25] in this case worked better than the correlation measures of Hidaka et al. (see Table 2).

For assessing how well the t_{mp50}-values are directly predicted by the signal-based parameters t_m (and not as part of a regression model), Fig. 8 displays all assessed parameters in the $t_m t_{mp50}$-plane. If predicted values of t_{mp50} were identical to the estimated mixing times t_{mp}, points should scatter along the angle bisector of the $t_m t_{mp50}$-plane. As can be seen, this is again best achieved by Abel and Huang's approach (criterion I, [25]).

4 CONCLUSIONS

The perceptual mixing time was assessed for the first time by means of a high quality dynamic binaural simulation. BRIR data sets have been acquired for nine acoustical environments, systematically varied in volume and average

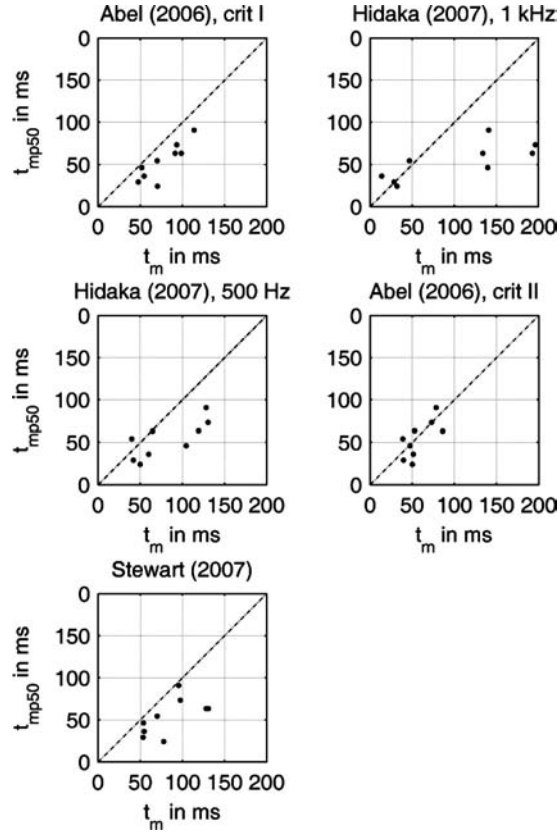

Fig. 8. Average perceptual mixing times t_{mp50} (mean values from both channels of a BRIR) plotted over signal-based t_m-predictors.

absorption. Both model- and signal-based estimators of the mixing time were evaluated for their power to predict the listening test results of a group of expert listeners. As a result, linear regression models predicting either (a) the average, or (b) the more critical 95%-point of the perceptual mixing times were presented, yielding predictors for situations, where (1) only the dimensions of the room are known (for instance in the case of model-based auralization), or (2) when an impulse response is available (for instance in the case of data-based auralization).

Results show that for shoebox shaped rooms the average perceptual mixing time can be well predicted by the enclosure's ratio of volume over surface area V/S (12) and by $\sqrt{V}$ (13) being indicators of the mean free path length, and the reflection density, respectively. The linear factors in our regression models suggest that a time interval corresponding to about two mean free path lengths, i.e., on average two orders of reflection, and a reflection density of less than 200 s^{-1} is perceived as diffuse even by trained listeners. Any dependence on reverberation time turned out to be due to its implicit covariation with room volume.

If an impulse response is available, average perceptual mixing times can be optimally predicted by regression formula (15) using values calculated from the echo density approach of Abel and Huang [25] applying the stopping criterion I (the echo density profile becoming equal to unity). For increased reliability of the prediction, the input value should be an average over several measurement positions.

The presented regression formulae for a perceptual mixing time can be applied to reduce the rendering effort of both loudspeaker- or headphone-based high quality VAEs and plausible auralization on limited platforms such as mobile audio devices.

For convenient application of the presented predictors we made appropriate Matlab® source code publicly available[1].

5 ACKNOWLEDGMENTS

Alexander Lindau was supported by a grant from the Deutsche Telekom Laboratories and by the Deutsche Forschungsgemeinschaft (DFG), grant WE 4057/1-1.

6 REFERENCES

[1] J.-D. Polack, "Modifying Chambers to Play Billiards the Foundations of Reverberation Theory," *Acustica*, vol. 76, pp. 257-272 (1992).

[2] A. Reilly, D. McGrath, "Real-Time Auralization with Head Tracking," *Proc. of the 5th Australian Regional AES Conv.*, Sydney (1995), preprint no. 4024.

[3] K. Meesawat, D. Hammershøi, "The Time when the Reverberant Tail in Binaural Room Impulse Response Begins," presented at *the 115th Convention of the Audio Engineering Society (2003 Oct.)*, convention paper 5859.

[4] F. Menzer, C. Faller, "Investigations on an Early-Reflection-Free Model for BRIRs," *J. Audio Eng. Soc.*, vol. 58, pp. 709-723 (2010 Sept.).

[5] S. E. Olive, F. E. Toole, "The Detection of Reflections in Typical Rooms," *J. Audio Eng. Soc.*, vol. 37, pp. 539-553 (1989 Jul./Aug.).

[6] A. Lindau, T. Hohn, S. Weinzierl, "Binaural Resynthesis for Comparative Studies of Acoustical Environments," presented at *the 122nd Convention of the Audio Engineering Society* (2007 May), convention paper 7032.

[7] H. Kuttruff, *Room Acoustics,* 4th ed. (New York : Routledge Chapman & Hall 2000).

[8] M.R. Schroeder "Measurement of Sound Diffusion in Reverberation Chambers," *J. Acoust. Soc. Am.*, vol. 31, no. 11, pp. 1407-1414 (1959).

[9] B. Blesser "An Interdisciplinary Synthesis of Reverberation Viewpoints," *J. Audio Eng. Soc.*, vol. 49, pp. 867-903 (2001 Oct.).

[10] W. B. Joyce "Sabine's Reverberation Time and Ergodic Auditoriums," *J. Acoust. Soc. Am.*, vol. 58, no. 3, pp. 643-655 (1975).

[11] D. Begault, "Perceptual Effects of Synthetic Reverberation on Three-Dimensional Audio Systems," *J. Audio Eng. Soc.*, vol. 40, pp. 895-904 (1992 Nov.).

[12] J. S. Bradley, G. A. Soulodre "The Influence of Late Arriving Energy on Spatial Impression," *J. Acoust. Soc. Am.*, vol. 97, no. 4, pp. 2263-2271 (1995).

[13] H. Kuttruff "Auralization of Impulse Responses Modeled on the Basis of Ray-Tracing Results," *J. Audio Eng. Soc.*, vol. 41, pp. 876-880 (1993 Nov.).

[14] W. Reichardt, U. Lehmann, "Raumeindruck als Oberbegriff von Räumlichkeit und Halligkeit, Erläuterungen des Raumeindrucksmaßes R," *Acustica*, vol. 40, no. 5, pp. 277-290 (1978).

[15] T. Hidaka, T. Okano, L. L. Beranek, "Interaural Cross-Correlation, Lateral Fraction, and Low- and High-Frequency Sound Levels as Measures of Acoustical Quality in Concert Halls," *J. Acoust. Soc. Am.*, vol. 98, no. 2, pp. 988-1007 (1995).

[16] L. Cremer, H. A. Müller, *Die wissenschaftlichen Grundlagen der Raumakustik. Bd. 1: Geometrische Raumakustik. Statistische Raumakustik. Psychologische Raumakustik*, 2nd ed. (Stuttgart : Hirzel, 1978).

[17] W. Schmidt, W. Ahnert, "Einfluss der Richtungs- und Zeitdiffusität von Anfangsreflexionen auf den Raumeindruck," *Wiss. Zeit. d. TU Dresden*, vol. 22, pp. 313 ff. (1973).

[18] J.-D. Polack, *La transmission de l'énergie sonore dans les salles*, Thèse de Doctorat d'Etat. Le Mans : Université du Maine (1988).

[19] M. R. Schroeder "Natural Sounding Artificial Reverberation," *J. Audio Eng. Soc.*, vol. 10, pp. 219-223 (1962 July).

[20] L. Schreiber, "Was empfinden wir als gleichförmiges Rauschen?" *Frequenz*, vol. 14, no. 12, pp. 399 ff. (1960).

[21] P. Rubak, L. G. Johansen, "Artificial Reverberation Based on a Pseudo-Random Impulse Response II," *presented at 106th Convention of* the Audio Engineering Society (1999 May), preprint 4900.

[22] D. Griesinger, "Practical Processors and Programs for Digital Reverberation," *Proceedings of the 7th International AES Conference: Audio in Digital Times (1989 May), paper 7-027.*

[23] T. Hidaka, Y. Yamada, T. Nakagawa "A New Definition of Boundary Point between Early Reflections and Late Reverberation in Room Impulse Responses," *J. Acoust. Soc. Am.*, vol. 122, no. 1, pp. 326-332 (2007).

[24] M. R. Schroeder, "New Method of Measuring Reverberation Time," *J. Acoust. Soc. Am.*, vol. 37, pp. 409-412 (1965).

[25] J. S. Abel, P. Huang "A Simple, Robust Measure of Reverberation Echo Density," presented at the *121st Convention of the Audio Engineering Society*(2006 Oct.), convention paper 6985.

[26] R. Stewart, M. Sandler "Statistical Measures of Early Reflections of Room Impulse Responses," *Proc. of the 10th Int. Conference on Digital Audio Effects (DAFx-07)*. Bordeaux (2007).

[27] G. Defrance, L. Daudet, J.-D. Polack "Using Matched Pursuit for Estimating Mixing Time Within Room Impulse Responses," *Acta Acustica united with Acustica*, vol. 95, no. 6, pp. 1071-1081 (2009).

[28] H. Kuttruff, "On the Audibility of Phase Distortion in Rooms and its Significance for Sound Reproduction and

[1] http://www.ak.tu-berlin.de/menue/digitale_ressourcen/research_tools/mixing_time_prediction

Digital Simulation in Room Acoustics," *Acustica*, vol. 74, pp. 3-7 (1991).

[29] D. de Vries, E. M. Hulsebos, and J. Baan, "Spatial Fluctuations in Measures for Spaciousness," *J. Acoust. Soc. Am.*, vol. 110, no. 2, pp. 947-954 (2001).

[30] Z. Schärer, A. Lindau, "Evaluation of Equalization Methods for Binaural Signals," presented at the *126th Convention of the Audio Engineering Society*(2009 May): , preprint 7721.

[31] A. Lindau, F. Brinkmann "Perceptual Evaluation of Headphone Compensation in Binaural Synthesis Based on Non-Individual Recordings," In: *J. Audio Eng. Soc.*, Vol. 60, No. 1/2, pp. 54-62 (2012).

[32] S. Ciba, A. Wlodarski, H.-J. Maempel "WhisPER – A New Tool for Performing Listening Tests," presented at the *126th Convention of the Audio Engineering Society* (2009 May), convention paper 7749.

[33] P. E. King-Smith, "Efficient and Unbiased Modifications of the QUEST Threshold Method: Theory, Simulations, Experimental Simulations, Experimental Evaluation and Practical Implementation," *Vision Research*, vol. 34, no. 7, pp. 885-912 (1994).

THE AUTHORS

A. Lindau

L. Kosanke

S. Weinzierl

Alexander Lindau was born 1976 in Berlin, Germany. In 2006 he received a M.A. degree in communication sciences, electrical engineering, and technical acoustics from the TU Berlin, Germany. From 2007–2010 he held a Ph.D. grant of the Deutsche Telekom Laboratories. Since 2011 he is a researcher in the research unit SEACEN (Simulation and Evaluation of Acoustical Environments), funded by the Deutsche Forschungsgemeinschaft. His fields of interest are room and electro acoustics, measurement techniques, binaural synthesis, and perceptual audio evaluation. Mr. Lindau is student member of AES, ASA, and DEGA (Germany).

●

Linda Kosanke was born in 1985 in Hennigsdorf, Germany. In 2011 she received an M.A. degree in communication sciences and technical acoustics from the TU Berlin, Germany. Since then she has been working for an engineering company and is primarily concerned with room acoustic design and simulation. Linda Kosanke is member of DEGA (Germany).

●

Stefan Weinzierl was born in 1967 in Bamberg, Germany. With a diploma in physics and sound recording (Tonmeister), a Ph.D. in systematic musicology, and after some years working as a recording producer for several major record labels, he joined the TU Berlin in 2004 as head of the Audio Communication Group. His fields of interest include audio technology, musical acoustics, room acoustics, and virtual acoustics. He is coordinating a master program in audio communication and technology at TU Berlin and he is coordinating a research consortium on virtual acoustics (SEACEN, DFG FOR 1557).