Validating a real-time perceptual model predicting distraction caused by audio-on-audio interference

Rämö, Jussi; Bech, Søren; Jensen, Søren Holdt

Published in:
The Journal of the Acoustical Society of America

DOI (link to publication from Publisher):
10.1121/1.5045321

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Rämö, J., Bech, S., & Jensen, S. H. (2018). Validating a real-time perceptual model predicting distraction caused by audio-on-audio interference. *The Journal of the Acoustical Society of America*, *144*(1), 153-163. https://doi.org/10.1121/1.5045321

Validating a real-time perceptual model predicting distraction caused by audio-on-audio interference

Jussi Rämö, Søren Bech, and Søren Holdt Jensen

Citation: The Journal of the Acoustical Society of America **144**, 153 (2018); doi: 10.1121/1.5045321
View online: https://doi.org/10.1121/1.5045321
View Table of Contents: http://asa.scitation.org/toc/jas/144/1
Published by the Acoustical Society of America

Validating a real-time perceptual model predicting distraction caused by audio-on-audio interference

Jussi Rämö,[1,a)] Søren Bech,[2,b)] and Søren Holdt Jensen[1]
[1]*Department of Electronic Systems, Aalborg University, Fredrik Bajers Vej 7, Aalborg, 9220, Denmark*
[2]*Bang & Olufsen a/s, Peter Bangs Vej 15, Struer, 7600, Denmark*

(Received 27 September 2017; revised 17 May 2018; accepted 16 June 2018; published online 9 July 2018)

This work concentrates on validating a real-time perceptual model predicting distraction caused by audio-on-audio interference. The real-time model was recently developed on the basis of another successfully validated, perceptual distraction model, which is not able to calculate predictions in real time. Both models are non-blind, i.e., their inputs take target and interferer signals separately. This paper describes a validation experiment for the real-time distraction model, which compares the model's predictions to subjective distraction ratings obtained from a listening experiment. The accuracy of the real-time model is also compared to that of the original distraction model. The calculated root-mean-squared errors for a speech zone and a music zone were 10.2% and 12.6% for the real-time model, respectively, compared to 11.3% and 11.5% for the original model. The results indicate that the real-time model is able to predict the distraction with similar accuracy as the original model, and thus, is a suitable tool for sound-zone evaluation. Furthermore, the real-time capability of the model is considered to be vital for certain applications, including the evaluation of adaptive sound zones. https://doi.org/10.1121/1.5045321

[MAH] Pages: 153–163

I. INTRODUCTION

Audio-on-audio interference is constantly present in our everyday lives, when two or more sound sources are competing for our attention. The sound sources can be natural sources or they can be audio systems, such as a TV, a radio, or portable speakers connected to a smartphone. This paper considers the latter case, where the sound sources are electrical devices, especially in the context of sound-zone systems.

The concept of sound zones was originally proposed by Druyvesteyn *et al.* (1994). In the past decade and a half, there have been plenty of contributions in developing different sound-zone concepts and methods, e.g., Betlehem *et al.*, 2015; Chang *et al.*, 2009; Choi and Kim, 2002; Møller *et al.*, 2012; Olik *et al.*, 2013; Pasco *et al.*, 2017; Schellekens and Møller, 2016; Shin *et al.*, 2010; Wu and Abhayapala, 2011; Zhu *et al.*, 2017. The main idea in sound zones is to create personal zones for multiple users within one acoustical space, where they are able to control and listen to their own audio content without disturbing other users.

In sound-zone scenarios, audio-on-audio interference occurs when the audio from other zones leak and interfere with the target audio that one is concentrating on in their own zone. Francombe *et al.* (2014a) conducted an elicitation experiment to define the perceptual attributes that describe the experience of audio-on-audio interference while listening to a target audio programme in the presence of an interfering audio programme. They concluded that *distraction* was by far the most important attribute for describing audio-on-audio interference, while *balance and blend* (of the two programme materials) came second.

Motivated by the elicitation study, Francombe *et al.* developed a perceptual model aimed at predicting the experienced distraction occurring in audio-on-audio interference situations (Francombe, 2014; Francombe and Baykaner, 2017; Francombe *et al.*, 2013, 2015). The model was originally trained by using a simple loudspeaker setup, consisting of only two loudspeakers, while at the same time considering that one of the main applications for the model would be the evaluation of sound-zone systems.

Recently, the performance of the model was successfully validated using two different sound-zone systems (Rämö *et al.*, 2016; Rämö *et al.*, 2017b). Although, the model is operating as expected in sound-zone environments, the main issue is that it is computationally slow, and thus, it is not possible to predict the distraction in real time, which would be beneficial and even obligatory for certain applications such as for adaptive sound zones.

Rämö *et al.* (2017a) have developed a modified version of the distraction model, which provides distraction predictions in real-time. This paper concentrates on validating the real-time distraction model by using a different sound-zone system, resulting in a different sound field, than during the development of the model.

The paper is organized as follows. Section II introduces both the original and the real-time distraction models. Section III describes the sound-zone setup and the recordings

[a)]Current address: Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland. Also at: Bang & Olufsen a/s, Peter Bangs Vej 15, Struer, 7600, Denmark. Electronic mail: jussi.ramo@aalto.fi
[b)]Also at Department of Electronic Systems, Aalborg University, Fredrik Bajers Vej 7, Aalborg, 9220, Denmark.

used in the validation procedure. Section IV presents the design and the results of the listening experiment used to obtain subjective distraction ratings. Section V shows the results for the model performance by comparing the model's predictions against the subjective distraction ratings. Finally, Sec. VI concludes this paper.

II. PERCEPTUAL DISTRACTION MODELS

This section describes the original distraction model implemented by Francombe *et al.* (Francombe, 2014; Francombe and Baykaner, 2017; Francombe *et al.*, 2013, 2015) and the modified version of that model, capable of real-time processing, developed by Rämö *et al.* (2017a).

A. Original distraction model

The distraction model (Francombe, 2014; Francombe and Baykaner, 2017; Francombe *et al.*, 2013, 2015) was trained to predict the distraction that users experience, when they are in an environment with two competing audio sources. The training setup consisted of two loudspeakers, one for the target audio, which the user was concentrating on, and one for the competing, interfering sound. Both the target and interferer stimuli were music (music-on-music) during the training of the model. Although, the training was conducted using such a simple loudspeaker setup, the model was aimed to be used in the context of sound zones.

The distraction model consists of five features and one constant term. The distraction prediction $\hat{y}$ is calculated from the features as follows:

$$\hat{y} = 24.19 + 1.04f_1 - 2.04f_2 - 0.41f_3 - 0.95f_4 - 0.16f_5, \quad (1)$$

where the features are described as

f_1: maximum long term loudness (LTL; Glasberg and Moore, 2002) when both the target and the interferer are active,

f_2: target-to-interferer ratio (TIR) using LTL,

f_3: interference-related perceptual score (IPS) from the PEASS software toolbox (Emiya *et al.*, 2011; Vincent, 2012),

f_4: the range of computational auditory signal-processing and perception (CASP) model (Jepsen *et al.*, 2008) output for the interferer signal at high frequencies (bands 20–31),

f_5: percentage of temporal windows (400 ms, 25% overlap) where TIR derived from the CASP model's outputs is less than 5 dB.

Furthermore, the model output $\hat{y}$ is limited between 0 (not at all distracting) and 100 (overpoweringly distracting). The model requires a dummy head recording as well as a typical single channel recording of the target and interferer signals from inside the zones.

B. Validation of the original model

The model has been previously validated in Rämö *et al.* (2016) and Rämö *et al.* (2017b) using two different complex sound-zone setups, illustrated in Fig. 1. In both setups, the idea is that zone A is for speech programmes, such as news or sport shows from TV or radio, and zone B is for music. Furthermore, the zones act as interfering sound sources to one another, so the target sound of zone A becomes the interferer in zone B and vice versa.

The first round of validations (Rämö *et al.*, 2016) were conducted using the setup depicted in Fig. 1(a). The validation experiment was carried out only in zone B, i.e., with music targets and speech interferers (speech-on-music). The aim of the second round of validations (Rämö *et al.*, 2017b) was to make physical alterations to the setup, as illustrated in Fig. 1(b), and validate the distraction model in both zones, i.e., with speech

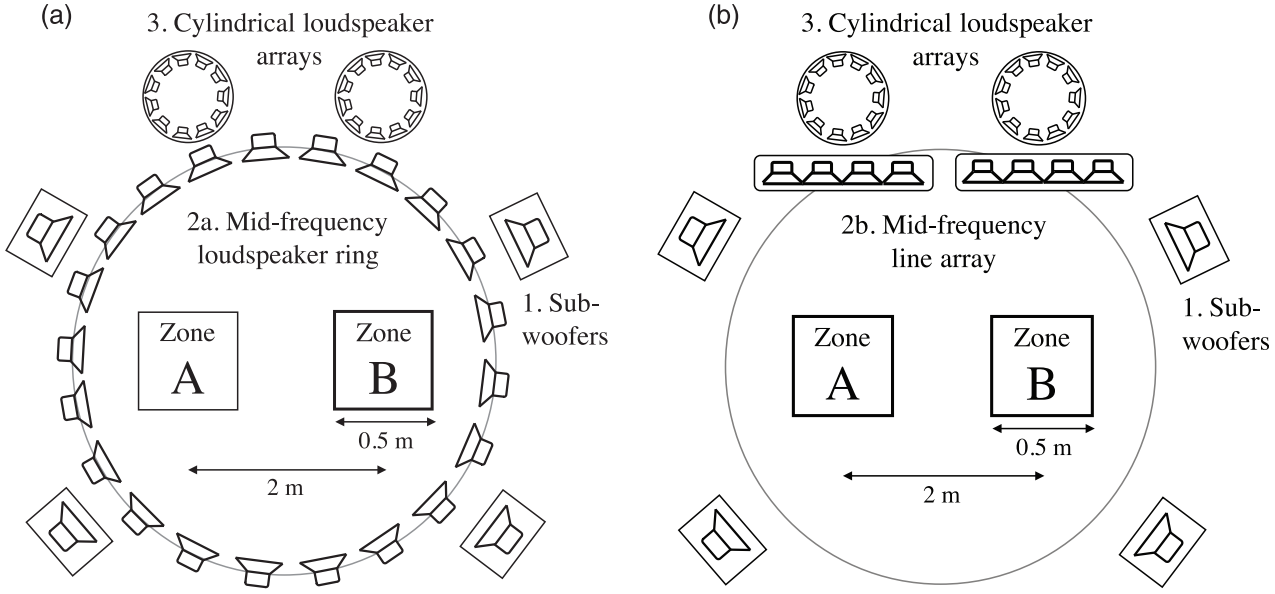

FIG. 1. Schematics of sound-zone setups where (a) illustrates the setup with a ring-shaped mid-frequency loudspeaker array used in the first validation experiment for the original model (Rämö *et al.*, 2016) and in the development of the real-time model in Rämö *et al.* (2017a), and (b) illustrates the setup with mid-frequency line arrays used in the second validation experiment of the original model (Rämö *et al.*, 2017b) and in the validation of the real-time distraction model in this paper. Note that the number of loudspeakers depicted does not represent the actual number of loudspeakers in these setups. Adapted from Rämö *et al.* (2017b).

TABLE I. Validation results for the original distraction model. Adapted from Rämö *et al.* (2017b).

Statistics	Training[a] Music-on-music	Validation I[b] Zone B	Validation II[c] Zone A	Validation II[c] Zone B
Root-mean-squared error (RMSE; %)	9.46	11.0	11.3	10.4
Epsilon-insensitive root-mean-squared-error (RMSE*; %)	4.41	5.56	4.48	5.09
R	0.94	0.99	0.98	0.99
R^2	0.88	0.96	0.95	0.95
Adjusted R^2	0.87	0.94	0.93	0.94

[a]Data from Francombe (2014).
[b]Data from Rämö *et al.* (2016).
[c]Data from Rämö *et al.* (2017b).

targets and music interferers (music-on-speech) in zone A and music targets and speech interferers in zone B.

Table I shows the validation results from both of the previous experiments (Rämö *et al.*, 2016; Rämö *et al.*, 2017b), as well as the results obtained from the training dataset that was used in developing the model (Francombe, 2014), i.e., the two-loudspeaker setup described in Sec. II A with music-on-music interference.

The root-mean-squared-error (RMSE) and the epsilon-insensitive root-mean-squared-error (RMSE*) describe the model's goodness of fit. Both the RMSE and RMSE* take the number of features used in the model into account, as described in Francombe (2014) and Rämö *et al.* (2017b), while the RMSE* also considers the subjective uncertainty in the data by utilizing the confidence intervals of the subjective data. The RMSE* considers there is an error between the predicted and observed values only when the predicted value lies outside the confidence interval, and then, the error is calculated as a distance between the predicted value and the nearest confidence interval bound.

Furthermore, correlation (R) measures the linear association between the observed and predicted values, whereas the R^2—the coefficient of determination—describes the amount of variance in the data explained by the model. Last, the adjusted R^2 is also aware of the number of features used in the predictive model, which is useful in order to avoid overfitting of the model by introducing too many features.

As can be seen in Table I, the validation results are close to that of the dataset that was used to train the model in Francombe (2014), suggesting that the model operates well when used in actual complex sound-zone environments (Rämö *et al.*, 2016; Rämö *et al.*, 2017b).

The main issue with the distraction model is that it is computationally slow, and thus, cannot operate in real time. The model takes approximately 13 min to calculate a single distraction prediction for a 10-s signal when using MATLAB and a Mid 2014 MacBook Pro (15-inch; Apple Inc., Cupertino, CA).

C. Real-time distraction model

The real-time distraction model (Rämö *et al.*, 2017a) was developed based on the original model. This was motivated by the successful validation experiments described in Sec. II B. The approach in developing the real-time distraction model was to utilize similar features as used in the original model [see Eq. (1) and the feature descriptions below that], but to replace the underlying algorithms with faster ones.

The main building block of the real-time model was chosen to be the ITU-R BS.1770-4 (2015) recommendation for multichannel loudness estimation. It was used to replace the Glasberg–Moore loudness algorithm and the CASP model, which were essential parts of the original distraction model estimating features f_1 and f_2, and f_4 and f_5, respectively. The PEASS algorithm used in the calculation of f_3 could not be directly replaced by using the ITU-loudness algorithm. However, the calculation of the real-time version of feature 3 is still indirectly based on the ITU-loudness algorithm, since it is estimated using the TIR calculated with the ITU-loudness algorithm for feature 2, as shown in Eq. (4). The equation is empirically derived, as explained in Rämö *et al.* (2017a), and it is not directly related to the PEASS algorithm.

Figure 2 shows the block diagram of the real-time distraction model where the input signals are dummy head recordings of the target audio, interfering audio, and their combination. Furthermore, the first four blocks inside the

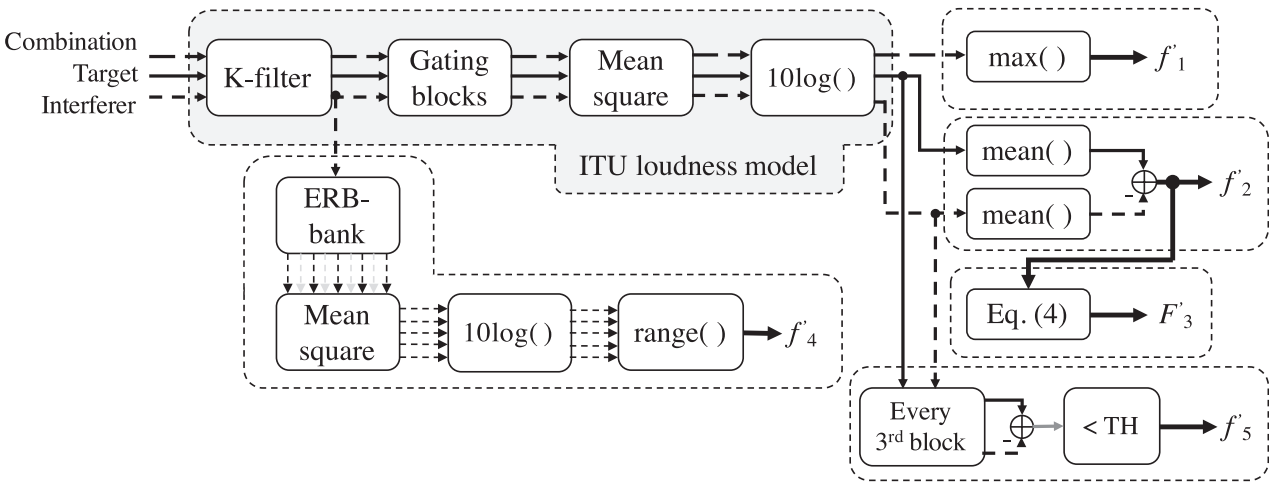

FIG. 2. Block diagram of the real-time distraction model based on the ITU recommendation ITU-R BS.1770-4 (2015). Function block 10 log( ) is described in Eq. (2). Adapted from Rämö *et al.* (2017a).

gray box depict the building blocks of the ITU-loudness algorithm, where the K-filtering block consists of two cascaded biquad filters, one accounting for the acoustics of the head and the other acting as a highpass filter. Only the highpass filter is included in the real-time distraction model since the acoustics of the head are physically taken into account by the dummy head. Moreover, the gating blocks are defined to be 400 ms with 75% overlap, after which the loudness of the jth gating block is calculated using

$$l_j = -0.691 + 10\log(z_j), \tag{2}$$

where z_j is the mean square of the jth gating block

The features of the real-time distraction model are illustrated in Fig. 2 and described as follows:

f'_1: maximum ITU-based loudness, when both the target and the interferer are active,

f'_2: TIR based on the ITU loudness estimation,

F'_3: calculated based on f'_2; see Eq. (4),

f'_4: the range of the ITU loudness estimation of the interferer signal at high frequencies [equivalent rectangular bandwidth (ERB) motivated bands 20–31],

f'_5: percentage of temporal windows (400 ms, 25% overlap) where TIR based on the ITU loudness estimations is less than 13 dB.

The distraction prediction $\hat{y}'$ is obtained from these features with

$$\hat{y}' = 24.19 + 1.04f'_1 - 2.04f'_2 + F'_3 - 0.95f'_4 - 0.16f'_5, \tag{3}$$

where F'_3 is

$$F'_3 = \begin{cases} 0, & \text{when} \quad f'_2 < 0 \\ 2.04f'_2, & \text{when} \quad 0 \le f'_2 \le 20 \\ -40, & \text{when} \quad f'_2 > 20 \end{cases}. \tag{4}$$

The model was trained and tested with the same dataset used in the first validation experiment of the original model (Rämö *et al.*, 2016), i.e., the speech-on-music stimuli in zone B reproduced with the sound-zone setup depicted in Fig. 1(a). The main idea in the training of the model was to fine-tune the model features ($f'_{1–5}$) to match those of the original model ($f_{1–5}$), so it would be possible to use the same model weights in the real-time model, in Eq. (3), as were used in the original model in Eq. (1).

The results showed that the real-time model had a RMSE of 10.9% compared to the RMSE of 11.0% for the original model's predictions of the same data (Rämö *et al.*, 2017a). Thus, the accuracy of the real-time model was similar to that of the original model, while only taking 0.04% of the computational time that was used by the original model. Another benefit of the real-time model is that it requires only the dummy head recordings and not the extra single channel recording needed in the original model.

III. SOUND-ZONE SETUP

This section describes the sound-zone setup, stimuli, and the recording procedure for the stimuli used in the validation experiment described in Secs. IV and V. The validation setup for the real-time model, illustrated in Fig. 1(b), was the same one that was used in the second round of validations of the original model in Rämö *et al.* (2017b).

The sound-zone setup was located in a large room (279 m^3) treated with absorption material, resulting in reverberation time of $T_{20–500\text{Hz}} < 0.6$ s and $T_{500–8000\text{Hz}} \approx 0.3$ s. The setup consisted of three loudspeaker arrays [see Fig. 1(b)]:

(1) Low-frequency array: 8 subwoofers,
(2) Mid-frequency array: 2 × 5-loudspeaker line arrays (one per zone),
(3) High-frequency array: 2 cylindrical loudspeaker arrays with 24 loudspeakers in each (one per zone).

The signal processing algorithm used to create the sound zones is a time-domain method broadband acoustic contrast control with pressure matching penalty (BACC-PM; Gálvez *et al.*, 2015; Møller and Olsen, 2016). The idea in BACC-PM is to minimize the reproduction error in the bright zone and the mean square pressure in the dark zone at the same time. The performance of the BACC-PM algorithm in the actual sound-zone setup [Fig. 1(b)] is illustrated in Figs. 3 and 4.

Figure 3 shows the measured acoustic contrast inside the zones where the two curves illustrate the difference between the target sound and the leaked interfering sound from the adjacent zone when both zones were playing white noise equally loud. The differences between the acoustic contrast plots in zones A and B are probably due to the room, e.g., zone B is located closer to a wall that is likely to reflect high frequencies more efficiently to zone B than to zone A.

By using the sound-zone setup with the BACC-PM algorithm, the TIR was on average 15 dB in both zones when calculated using the LA_{eq} (10 s) values from the target and interferer recordings. For example, if both targets in both zones are reproduced at the level of 70 dB, as illustrated in Fig. 4(a), both interfering audio programmes (i.e., the leaked sound from the adjacent zone) are attenuated by 15 dB, resulting in interferer levels of 55 dB in both zones. This means the TIR in both zones would be 75 dB – 55 dB = 15 dB.

The targets in zone A were speech signals and the targets in zone B were music signals. Moreover, since the zones acted as interfering sound sources to each other, the interferer in zone A was the leaked music signal from zone B, and the interferer in zone B was the leaked speech signal from zone A.

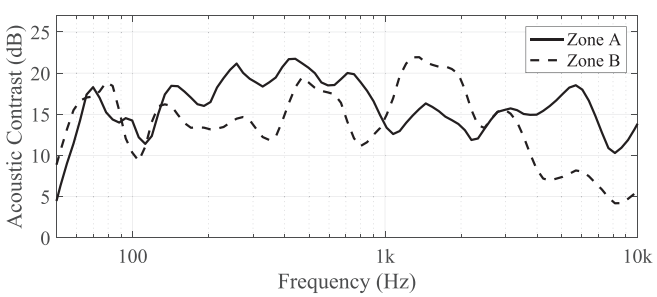

FIG. 3. Acoustic contrast within the zones measured using white noise and plotted with 1/3-octave smoothing.

A. Sound-zone recordings

The stimuli set used in this validation experiment was the same set that was used in Rämö *et al.* (2017b) to validate the original model. The stimuli set was recorded by placing a Brüel and Kjær (B&K, Nærum, Denmark) head-and-torso simulator (HATS) model 4100 in the sound-zone setup. The HATS was placed in both zones, one after another, and the target and interferer signals in both zones were recorded separately. The binaural HATS recordings were used as stimuli for the listening test as well as input signals for the real-time distraction model.

The programme materials reproduced with the sound-zone setup were the same loudness-matched speech and music samples used in Rämö *et al.* (2016). The samples were originally acquired from radio shows in the United Kingdom by a random radio sampling procedure described in Francombe *et al.* (2014b). The recorded speech and music samples were paired to form target-interferer pairs for both zones. This study used the same randomly chosen pairing as in Rämö *et al.* (2016) and Rämö *et al.* (2017b) in order to allow direct comparisons to previous results.

IV. LISTENING EXPERIMENT

This section describes a listening experiment conducted in order to obtain experimental data to be compared with the predictions of the real-time distraction model. The listening experiment has been previously reported and data from it have been used to validate the original distraction model in Rämö *et al.* (2017b). Using data derived from the same listening experiment as before gives us the possibility to compare the performance of the real-time model to that of the original model in Sec. V.

A. Stimuli

The stimuli set was built by having 31 music and speech sample pairs loudness matched to produce the maximum LTL of 70 dB in the their target zone (Rämö *et al.*, 2016). The first sample pair had a zero gain for both samples (i.e., both samples were equally loud in their target zone). After that, a gain was applied to one of the samples so that, first, the level of the target stimuli were kept constant at 70 dB and the level of the interferer stimuli were reduced by between 2 and 30 dB with 2 dB increments. This was also done the other way around, where the levels of the interferer stimuli were kept constant while the levels of the target stimuli were reduced. This way, every stimuli pair had a different TIR in the range from −15 to +45 dB with the increment of 2 dB. Note that the TIR is also different depending in which zone you are listening to the stimuli pair. For example, if the speech target is reproduced at 70 dB and the music target at 50 dB, the TIR in zone A would be 35 dB and in zone B it would be −5 dB, as illustrated in Fig. 4(b).

In addition to the 31 sample pairs, 5 reference stimuli were included in the experiment. The references had only the target sound and no interferer whatsoever. The references were added in order to check whether the participants had understood the task at hand correctly or not. Furthermore,

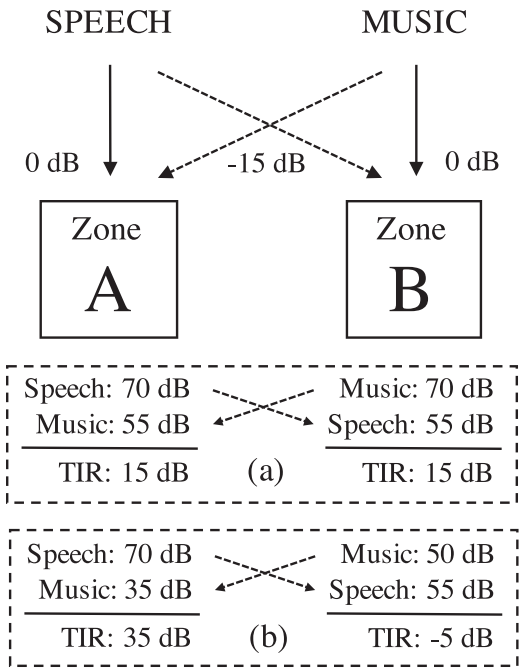

FIG. 4. Example TIRs gained from the sound-zone signal processing in the validation setup [Fig. 1(b)]. At the top, the solid lines depict the target sound (with normalized level at 0 dB) and the dashed lines illustrate the leaked interfering sound (attenuated by 15 dB compared to the target level). The example scenarios (a) and (b) illustrate how the level of the target programmes affect the TIRs in the two zones.

four of the normal sample pairs were randomly chosen to be repeated during the test (same samples for all subjects) in order to see how consistently the participants were able to rate them. In total, this resulted in 40 stimuli pairs that the participants were asked to evaluate in both zones.

B. Design

The listening test was conducted by using a pair of Sennheiser HD 600 headphones (Wedemark, Germany) to reproduce the above described stimuli set consisting of binaural HATS recordings with varying TIRs. It is generally known that headphone reproduction might have some issues while reproducing spatial aspects of recorded sound fields. However, the perceptual attributes used to evaluate audio-on-audio interference, like distraction evaluated in this paper, are not spatial by nature (Francombe *et al.*, 2014a).

The main motivations, on the other hand, for using headphone reproduction in our listening experiments in the first place were to avoid visual bias and enable scalability since we have experimented with two different sound-zone setups (so far), and practicality, allowing us to conduct listening experiments in multiple geographical locations (Rämö *et al.*, 2016; Rämö *et al.*, 2017b) without building multiple replicas of the quite complex sound-zone setups.

The listening level of the headphones was adjusted so the target stimuli, without any gain reductions applied to them, were reproduced at 70 dB with A-weighting. The level and the calibration procedure were the same as used before in Rämö *et al.* (2016), where the headphone level was adjusted using a B&K HATS (model 4128), including its diffuse field correction.

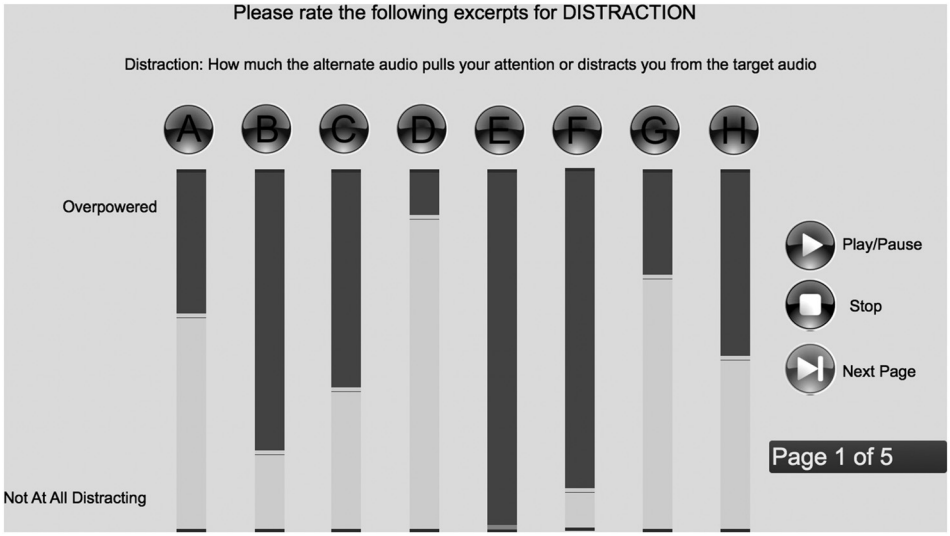

FIG. 5. GUI for the listening experiment, implemented with Max/MSP.

In the listening experiment, the participants were asked to rate the distraction they were experiencing due to the interfering sound leaking from the adjacent zone. The graphical user interface[1] (GUI), shown in Fig. 5, was a modified version of the interface used in Francombe (2014) and Rämö *et al.* (2016). The GUI defined distraction as "how much the alternate audio pulls your attention or distracts you from the target audio." The participants were presented eight stimuli per page with a slider for each stimulus. The sliders were unmarked except for the endpoint labels "not at all distracting" and "overpowered." The endpoint labels were located at values 10 and 90, while the full scale for the sliders was from 0 to 100, which is the same scale as used by the perceptual models. There were five pages in total, and each page contained a hidden reference—a target sound without any interferer.

Twenty-five persons participated in the listening experiment. They were aged between 21 and 60 yr. Twenty of them conducted the experiment at Aalborg University in Aalborg, Denmark, while the remaining five participants underwent the experiment at Bang & Olufsen premises in Struer, Denmark. Both groups used the exact same equipment during the listening experiment. There were both naive and more experienced listeners in the group of participants.

The hearing of each participant was tested with an audiometry test. No one was excluded from the experiment because of hearing loss, while the thresholds for exclusion were defined as a moderate hearing loss in one of the ears (45 dB hearing level or more) or mild hearing loss in both ears (25 dB hearing level or more).

C. Results

Figure 6 shows box plots of the listening experiment results for both zones, zone A in Fig. 6(a) and zone B in Fig. 6(b). The horizontal line inside each box shows the median of the distraction estimates for each sample pair, while the bottom of the box shows the 25th percentile and the top of the box shows the 75th percentile of the data. Furthermore, the whiskers extend a maximum of 1.5 times the height of the boxes, whereas the plus (+) markers depict outliers that do not fit within the whiskers. The numbers next to each outlier show the number of the participant in question. Furthermore, the five cases on the right-hand side of the vertical dashed line, in both Figs. 6(a) and 6(b), are the references, where there are no interferers present, and thus, the TIRs in these cases are marked as infinite. The horizontal dashed lines depict the distraction values of 10 and 90, where the endpoint labels were located in the GUI of the listening experiment; see Fig. 5.

The x axis shows the TIR of the stimuli inside the zone where the effect of the sound-zone signal processing is taken into account, as described in Sec. IV A and illustrated in Fig. 4. That is, when the TIR is 15 dB, both of the zones are playing their targets equally loud. With TIRs $>$ 15 dB, it means that the target level was kept constant at 70 dB and the level of the interferer was decreased, and with TIRs $<$ 15 dB, it is the opposite, i.e., the interferer level was kept constant while the target level was decreased.

When TIR is zero, the levels of the target sound and the interferer sound inside a zone are the same, which can be compared to a scenario where one would have two loudspeakers, one for the target and one for the interferer, playing equally loud without any sound-zone signal processing. Furthermore, as can be seen in Figs. 6(a) and 6(b), when TIR $<$ 0 dB, almost all subjects rated the scenarios to be distracting ($>$50), while the median values were over 80, i.e., highly distracting.

As the TIR increased, the distraction ratings decreased, as expected, since the target audio became more prominent. When TIRs were approximately between 0 and 28 dB, there was a lot of variance in the subjects' distraction ratings (i.e., tall boxes), probably because different people actually have quite a different standpoint as to what is distracting and what is not. However, after the TIR was above approximately 25 dB in zone A and 29 dB in zone B, the variance of the ratings plummeted and the participants seemed to have reached a consensus that the scenarios were not distracting anymore.

Before comparing the listening test results to the predictions of the model, a prescreening of the data was conducted

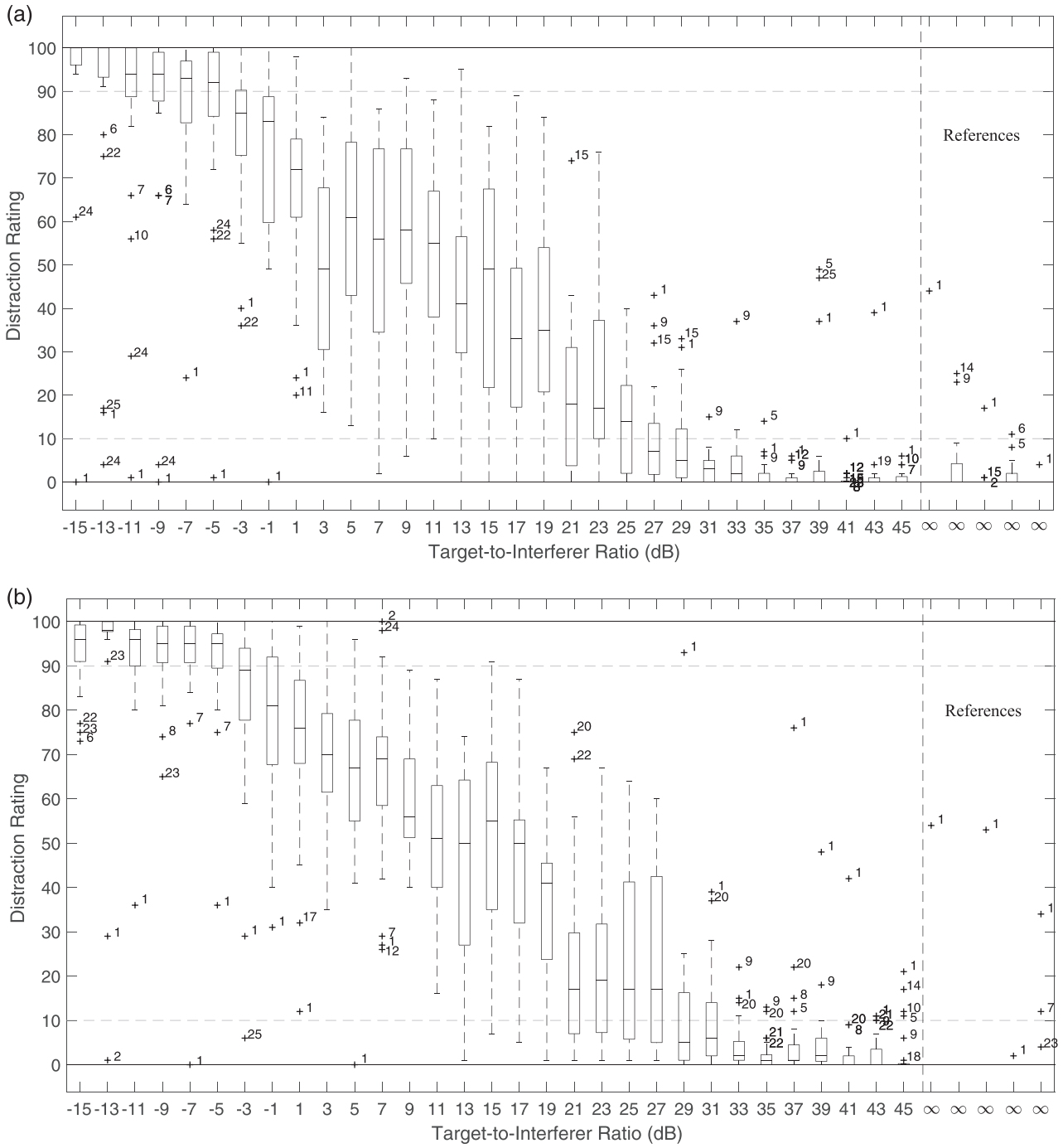

FIG. 6. Results of the listening experiment where (a) shows a box plot of zone A ratings and (b) shows a box plot of zone B ratings. The plus (+) markers depict outliers and the numbers next to them show the number of the participant in question.

in order to see whether the participants had correctly understood the task at hand or not, and if they had, could they provide consistent distraction ratings to the stimuli. The prescreening was conducted based on the hidden reference stimuli and the repeated stimuli that were included in the listening experiment.

Before the actual prescreening of the data, it can be clearly seen in Figs. 6(a) and 6(b) that participant no. 1 had somehow misunderstood the task, since they have 17 outliers in zone A ratings and 20 in zone B ratings, most of them at the opposing end of the distraction scale compared to the general opinion of the other participants. On top of that, participant no. 1 failed both repeatability and reliability tests in both zones, and thus, was excluded from all further analysis, including the prescreening of the references and repeated stimuli, discussed in Secs. IV D and IV E.

D. Prescreening: References

The threshold for the reference ratings was defined so that all of the references must be rated at "not at all distracting." In practice, all references had to be rated to be less than or equal to 10 on the distraction scale, i.e., where the "not at all distracting" label was placed in the GUI. Most of

the references were correctly rated to be near 0 by all participants, as can be seen in Figs. 6(a) and 6(b), however, there were two reference stimuli in zone A and one stimulus in zone B that were rated incorrectly by one or more participants (excluding participant no. 1, as explained above).

In zone A, there were three subjects that failed to rate one of the references below 10 on the distraction scale. Participant no. 6 gave the rating of 11 for the fourth reference (from the left) in Fig. 6(a), and thus, was excluded from further analysis based on the reference threshold. The ratings of participants nos. 9 and 14 for the second reference (from the left) raised some suspicions, since these two participants who were able to identify all other references correctly gave ratings as high as 23 and 25 for this one reference stimulus. It turned out that the target speech sample, recorded from a radio show, had also music in the background, starting after 5 s from the start of the sample. We suspect that the two participants mistook the music that was part of the target radio programme as being an interfering programme from the adjacent zone. It was decided not to exclude either of the participants from further analysis due to the ambiguous reference stimulus.

In zone B, there was only one participant (no. 7) who failed to give a correct rating to one of the reference stimuli [see Fig. 6(b)] and had to be excluded from the final analysis. All in all, the music references in zone B were consistently rated to 0 (slightly better than the speech references in zone A), as can be seen from the low variance in the reference ratings in Fig. 6(b).

E. Prescreening: Repeats

Figure 7 illustrates the ratings of the repeated stimuli used to evaluate the participants' consistency when rating the stimuli. The gray bars show the results in zone A and the black bars show the results in zone B. Each bar illustrates the mean of the absolute difference of a subject's ratings for the four repeated stimuli that were presented during the experiment. The gray and black solid lines show the overall mean across the subjects for zones A and B, respectively, while the dashed lines show the mean plus one standard deviation (SD), which was also used as a threshold for the repeatability screening, similarly as in Francombe (2014). Note that participant no. 1 is not included in Fig. 7 (as explained in Sec. IV C).

As can be seen in Fig. 7, for zone A there were three bars (participants) that were above the gray dashed line, illustrating the threshold. Thus, participant nos. 14, 16, and 25 were excluded from the final analysis of the zone A results. For zone B, there were five black bars above the threshold (black dashed line), namely, participant nos. 8, 11, 16, 20, and 22, and thus, they were also excluded from further analysis of the results.

V. MODEL PERFORMANCE

To summarize the prescreening results, there were four participants excluded from further analysis in zone A and six participants excluded from zone B, based on the evaluation of their ratings of the reference stimuli and the repeated stimuli. On top of that, participant no. 1 was removed from both zones, thus, leaving 20 valid participants in zone A and 18 in zone B. Furthermore, the references and repeated stimuli were removed from the dataset used to report the performance of the model in this section.

Figure 8 plots the prescreened data from the listening experiment as well as the predictions of the original distraction model [Figs. 8(a) and 8(b)] and the real-time distraction model [Figs. 8(c) and 8(d)] for both zones. The cross markers (×) depict the means of the experimental data accompanied with 95% confidence intervals calculated from the *t*-distribution, while the round markers (○) show the models' predicted values. The vertical dashed line in the middle shows the point where TIR is 15 dB and both zones are playing equally loud.

It is worth mentioning that although the experimental data used to validate the performance of the real-time model in this section are derived from the same listening experiment as utilized in Rämö *et al.* (2017b), the experimental results for zone B are somewhat different than reported in Rämö *et al.* (2017b). This is due to a slightly different prescreening procedure resulting in the exclusion of four more participants from zone B results than in Rämö *et al.* (2017b). The effect can be seen when comparing the results of the original model in zone B in Table I (last column) and Table II (second column).

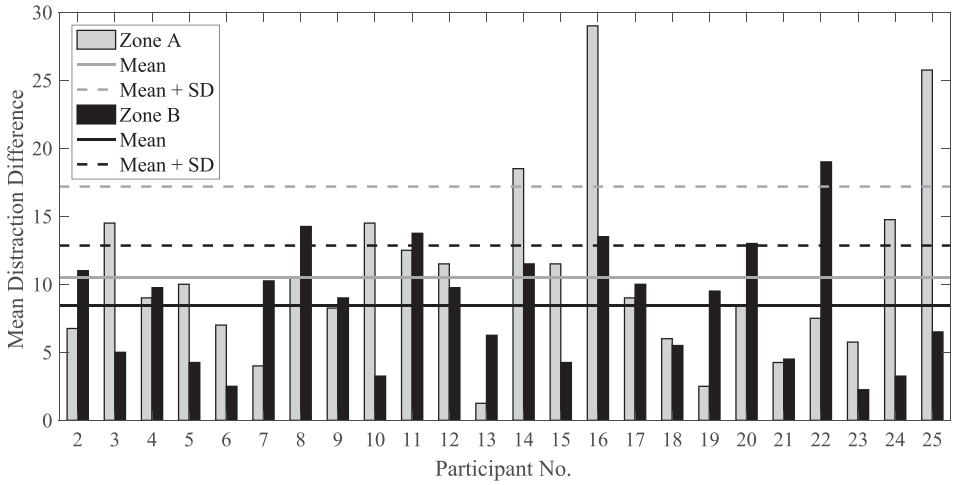

FIG. 7. Repeatability of the selected stimuli. The two solid lines show the overall mean and the dashed lines show the repeatability threshold—mean plus one standard deviation (SD)—for all participants in zone A (gray) and zone B (black).

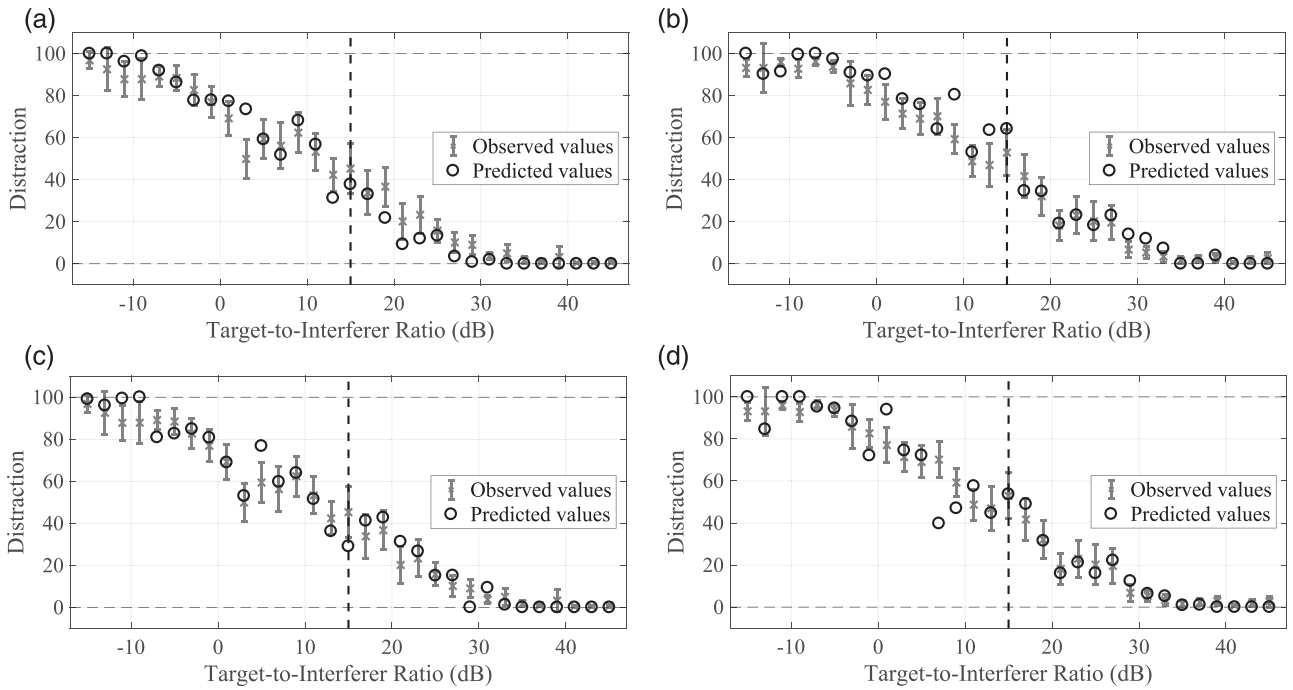

FIG. 8. Predictions of the distraction models compared against the subjective data from the listening experiment. The cross markers (×) show the mean of the subjective distraction ratings and the error bars indicate 95% confidence intervals. The round markers (○) show the predicted values of the distraction models, (a) and (b) show the predictions of the original model and (c) and (d) show the predictions of the real-time model. The vertical dashed line, in each figure, indicates the point where both zones are playing equally loud, i.e., TIR = 15 dB. (a) and (b) are adapted from Rämö *et al.* (2017b).

As can be seen in Figs. 8(c) and 8(d), the predictions of the real-time model were quite accurate when compared to the experimental results. In zone A results [Fig. 8(c)] there was only one clear misprediction when the TIR was 5 dB, while in zone B [Fig. 8(d)] there were two predictions that were clearly off, one overestimation and one underestimation, at TIR values of 1 dB and 7 dB, respectively.

The fact that the above-mentioned poorer predictions occurred at TIRs less than 10 dB is actually not so critical in practice as it would be with larger TIRs since it does not make a large difference what the actual distraction rating is, if it is distracting in any case. However, underestimations can be detrimental also in practice, i.e., when the model would suggest a certain scenario is not distracting, while actually it is perceived to be distracting, as was the case in Fig. 8(d) at TIR = 7 dB.

Figure 9 plots the predicted distraction against the observed distraction for both zones and both models. Figures 9(a)–9(d) illustrate the balance of over- and underpredictions in different cases. The cross markers that are above the $y = x$ dashed line are overpredictions, and the markers that are below the dashed line are underpredictions. For example, Fig. 9(c) shows that the real-time model in zone A operates well with only a few underpredictions. Moreover, two of those were in the range where the observed distraction was around 90, which, as mentioned before, does not matter in practical applications.

TABLE II. Validation results of the real-time model, compared against that of the original model.

	Original model		Real-time model	
Statistics	Zone A	Zone B	Zone A	Zone B
RMSE (%)	11.3	11.5	10.2	12.6
RMSE* (%)	4.48	5.21	3.29	7.23
R	0.98	0.99	0.98	0.98
R^2	0.95	0.96	0.96	0.95
Adjusted R^2	0.93	0.94	0.94	0.93

Table II shows the results in the form of statistics calculated based on the experimental data and predicted data, shown in Figs. 8(a)–8(c). The statistics in Table II show that the real-time model's predictions were slightly more accurate in zone A (RMSE = 10.2%) than in zone B (RMSE = 12.6%), which is a good result, while at the same time a bit surprising since both the original model and the real-time model were trained using music targets, corresponding more to zone B than to zone A.

When TIR was approximately 10 dB or more, the predictions of the real-time model were accurate in both zones, i.e., almost all predictions were within the confidence intervals, as shown in Figs. 8(c) and 8(d). Furthermore, when the accuracy was evaluated within the TIR range from 0 to 20 dB—the range where current sound-zone systems operate—the RMSE values were 7.2% and 11.2% for zones A and B, respectively. These values are slightly better compared to the full-range RMSE values shown in Table II, indicating the model's capability of providing accurate distraction predictions in the general sound-zone operating range.

As a result, the fact that the model was able to operate well in both zones is rather significant since it suggests that a single model can be used to evaluate the whole sound-zone system, consisting of both speech and music programmes.

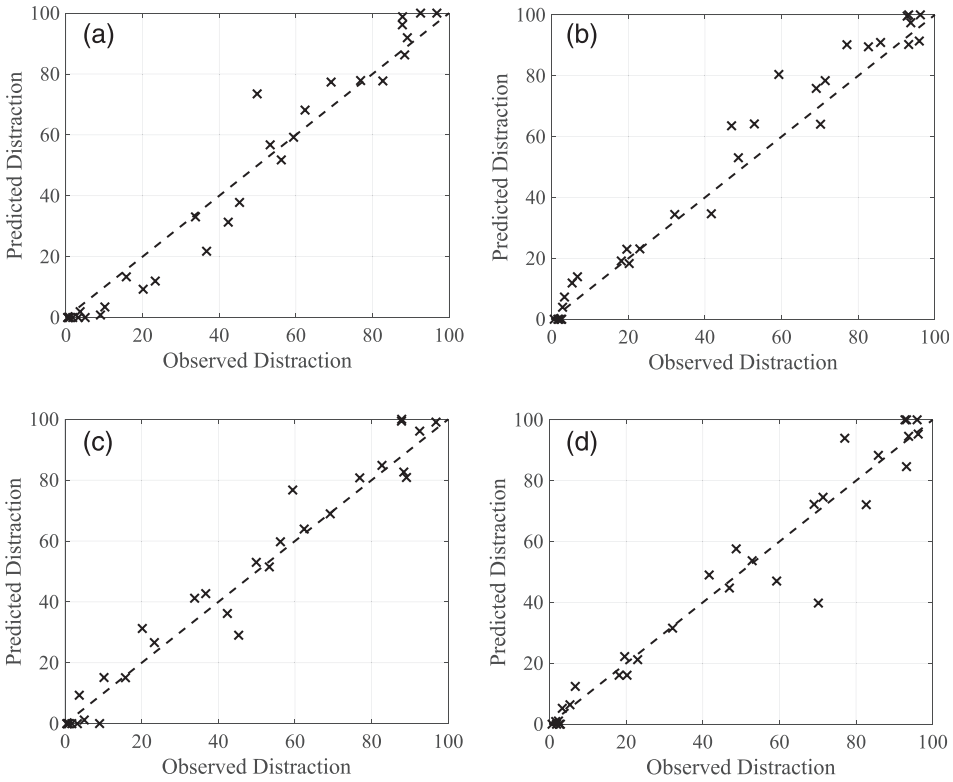

FIG. 9. Predictions of the distraction models plotted against the subjective observations. The diagonal dashed line in each figure indicates the $y = x$ line, where all the $\times$ markers would lie in an ideal situation. (a) and (b) are adapted from Rämö *et al.* (2017b).

Furthermore, the prediction accuracy is comparable to that of the original model, as shown in Figs. 8 and 9, and in Table II, suggesting that the real-time model can be used to predict user-experienced distraction instead of the much slower original model.

VI. CONCLUSION

This paper concentrated on validating a real-time perceptual model that predicts the distraction experienced by users in audio-on-audio interference situations. The real-time model is based on a previous distraction model, which has been validated in two different sound-zone setups in order to ensure that the model works correctly in real-life audio-on-audio interference scenarios. The real-time model estimates similar features as the original model, but utilizes different, faster algorithms to obtain these features, the main one being the ITU-R BS.1770-4 (2015) loudness estimation algorithm.

A listening experiment was conducted in order to get subjective data on different situations within a sound-zone environment, including different TIRs and different music and speech targets and interferers. The data from the listening experiment were then compared against the predictions of the real-time model, and the performance was also compared to that of the original model.

The results of the real-time model are highly comparable to those of the original model, indicating that the real-time model is able to provide accurate predictions. The RMSEs for the real-time model were 10.2% and 12.6% for zones A and B, respectively, while the corresponding RMSEs for the original model were 11.3% and 11.5%, respectively.

The main benefit of the real-time model over the original distraction model is its improvement in computational speed. For the original model, it takes approximately 13 min to calculate a single distraction prediction for a 10-s sample, while the real-time model is able to produce a prediction in 0.3 s. In practice, this means that the real-time model can be used to monitor the performance of a sound-zone system continuously while the sound-zone setup is active, which was not possible with the original distraction model.

The real-time distraction model can be used as a tool when designing, evaluating, and monitoring sound-zone systems. Furthermore, one major benefit of the real-time capability of the model is that the distraction predictions could be used as a control parameter for adaptive sound-zone systems, where the performance of the system can be optimised in real time, based on, e.g., current programme materials, number or location of users, or available loudspeaker configurations. Moreover, the use of the model is not limited in sound-zone environments, but could also be utilized in other applications where audio-on-audio interference is present.

ACKNOWLEDGMENTS

Parts of this work have been previously published and presented at the 173rd Meeting of the Acoustical Society of America and 8th Forum Acousticum, in Boston, MA, June 2017 (Rämö *et al.*, 2017b). These published portions are the listening test, providing the raw subjective data for this work, and the results of the original distraction model. Furthermore, the detailed description of the real-time distraction model was introduced in Rämö *et al.* (2017a). This work has been partly funded by the Innovation Fund Denmark (Project No. 4228-00002B). The authors would like to thank Mr. Martin Møller for his help and expertise on everything related to sound zones.

[1]Originally based on a Max/MSP patcher for MUSHRA listening tests, available from https://github.com/IoSR-Surrey/MUSHRA-MaxMSP (Last viewed 28 June 2018).

Betlehem, T., Zang, W., Poletti, M. A., and Abhayapala, T. D. (**2015**). "Personal sound zones," IEEE Signal Process. Mag. **32**(2), 81–91.

Chang, J.-H., Lee, C.-H., Park, J.-Y., and Kim, Y.-H. (**2009**). "A realization of sound focused personal audio system using acoustic contrast control," J. Acoust. Soc. Am. **125**(4), 2091–2097.

Choi, J.-W., and Kim, Y.-H. (**2002**). "Generation of an acoustically bright zone with an illuminated region using multiple sources," J. Acoust. Soc. Am. **111**(4), 1695–1700.

Druyvesteyn, W. F., Aarts, R. M., Asbury, A. J., Gelat, P., and Ruxton, A. (**1994**). "Personal sound," in *Proc. Inst. Acoust.*, Vol. 16, pp. 571–585.

Emiya, V., Vincent, E., Harlander, N., and Hohmann, V. (**2011**). "Subjective and objective quality assessment of audio source separation," IEEE Trans. Audio, Speech, Lang. Process. **19**(7), 2046–2057.

Francombe, J. (**2014**). "Perceptual evaluation of audio-on-audio interference in a personal sound zone system," Ph.D. thesis, University of Surrey, Guildford, UK.

Francombe, J., and Baykaner, K. (**2017**). "System and a method of providing sound to two sound zones" U.S. Patent US 9,635,483 B2.

Francombe, J., Mason, R., Dewhirst, M., and Bech, S. (**2013**). "Modeling listener distraction resulting from audio-on-audio interference," in *Proc. Mtgs. Acoust.*, Montreal, Canada, Vol. 19, pp. 1–9.

Francombe, J., Mason, R., Dewhirst, M., and Bech, S. (**2014a**). "Elicitation of attributes for the evaluation of audio-on-audio interference," J. Acoust. Soc. Am. **136**(5), 2630–2641.

Francombe, J., Mason, R., Dewhirst, M., and Bech, S. (**2014b**). "Investigation of a random radio sampling method for selection ecologically valid music programme material," in *Proc. AES 136th Conv.*, Berlin, Germany, pp. 1–10.

Francombe, J., Mason, R., Dewhirst, M., and Bech, S. (**2015**). "A model of distraction in an audio-on-audio interference situation with music program material," J. Audio Eng. Soc. **63**(1/2), 63–77.

Gálvez, M. F. S., Elliott, S. J., and Cheer, J. (**2015**). "Time domain optimization of filters used in a loudspeaker array for personal audio," IEEE/ACM Trans. Audio, Speech, Lang. Process. **23**(11), 1869–1878.

Glasberg, B. R., and Moore, B. C. J. (**2002**). "A model of loudness applicable to time-varying sounds," J. Audio Eng. Soc. **50**(5), 331–342.

ITU-R BS. 1770-4 (**2015**). "Algorithms to measure audio programme loudness and true-peak audio level" (International Telecommunication Union, Geneva, Switzerland).

Jepsen, M. L., Ewert, S. D., and Dau, T. (**2008**). "A computational model of human auditory signal processing and perception," J. Acoust. Soc. Am. **124**(1), 422–438.

Møller, M., and Olsen, M. (**2016**). "Sound zones: On performance prediction of contrast control methods," in *Proc. AES Int. Conf.*, Guildford, UK, pp. 1–10.

Møller, M., Olsen, M., and Jacobsen, F. (**2012**). "A hybrid method combining synthesis of a sound field and control of acoustic contrast," in *Proc. AES 132nd Conv.*, Budapest, Hungary, pp. 1–8.

Olik, M., Francombe, J., Coleman, P., Jackson, P. J. B., Olsen, M., Møller, M., Mason, R., and Bech, S. (**2013**). "A comparative performance study of sound zoning methods in a reflective environment," in *Proc. AES 52nd Int. Conf.*, Guildford, UK, pp. 1–10.

Pasco, Y., Gauthier, P.-A., Berry, A., and Moreau, S. (**2017**). "Interior sound field control using generalized singular value decomposition in the frequency domain," J. Acoust. Soc. Am. **141**(1), 334–345.

Rämö, J., Bech, S., and Jensen, S. H. (**2017a**). "Real-time perceptual model for distraction in interfering audio-on-audio scenarios," IEEE Signal Process. Lett. **24**(10), 1448–1452.

Rämö, J., Christensen, L., Bech, S., and Jensen, S. H. (**2017b**). "Validating a perceptual distraction model using a personal two-zone sound system," in *Proc. Mtgs. Acoust.*, Vol. 30, pp. 1–12.

Rämö, J., Marsh, S., Bech, S., Mason, R., and Jensen, S. H. (**2016**). "Validation of a perceptual distraction model in a complex personal sound zone system," in *Proc. AES 141st Conv.*, Los Angeles, CA, pp. 1–10.

Schellekens, D. H. M., and Møller, M. (**2016**). "Time domain acoustic contrast control implementation of sound zones for low-frequency input signals," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, Shanghai, China, pp. 365–369.

Shin, M., Lee, S. Q., Fazi, F. M., Nelson, P. A., Kim, D., Wang, S., Park, K. H., and Seo, J. (**2010**). "Maximization of acoustic energy difference between two spaces," J. Acoust. Soc. Am. **128**(1), 121–131.

Vincent, E. (**2012**). "Improved perceptual metrics for the evaluation of audio source separation," in *10th Int. Conf. Latent Variable Anal. and Signal Separation (LVA/ICA)*, Tel Aviv, Israel, pp. 430–437.

Wu, Y. J., and Abhayapala, T. D. (**2011**). "Spatial multizone soundfield reproduction: Theory and design," IEEE Trans. Audio, Speech, Lang. Process. **19**(6), 1711–1720.

Zhu, Q., Coleman, P., Wu, M., and Yang, J. (**2017**). "Robust acoustic contrast control with reduced *in-situ* measurement by acoustic modeling," J. Audio Eng. Soc. **65**(6), 460–473.