Preserving spatial perception in rooms using direct-sound driven dynamic range compression

Hassager, Henrik Gert; May, Tobias; Wiinberg, Alan; Dau, Torsten

Published in:
Journal of the Acoustical Society of America

Link to article, DOI:
10.1121/1.4984040

Publication date:
2017

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

Citation (APA):
Hassager, H. G., May, T., Wiinberg, A., & Dau, T. (2017). Preserving spatial perception in rooms using direct-sound driven dynamic range compression. Journal of the Acoustical Society of America, 141(6), 4556–4566. DOI: 10.1121/1.4984040

Preserving spatial perception in rooms using direct-sound driven dynamic range compression

Henrik Gert Hassager, Tobias May, Alan Wiinberg, and Torsten Dau[a)]

Hearing Systems Group, Department of Electrical Engineering, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark

(Received 27 January 2017; revised 29 April 2017; accepted 10 May 2017; published online 19 June 2017)

Fast-acting hearing-aid compression systems typically distort the auditory cues involved in the spatial perception of sounds in rooms by enhancing low-level reverberant energy portions of the sound relative to the direct sound. The present study investigated the benefit of a direct-sound driven compression system that adaptively selects appropriate time constants to preserve the listener's spatial impression. Specifically, fast-acting compression was maintained for time-frequency units dominated by the direct sound while the processing of the compressor was linearized for time-frequency units dominated by reverberation. This compression scheme was evaluated with normal-hearing listeners who indicated their perceived location and distribution of sound images in the horizontal plane for virtualized speech. The experimental results confirmed that both independent compression at each ear and linked compression across ears resulted in broader, sometimes internalized, sound images as well as image splits. In contrast, the linked direct-sound driven compression system provided the listeners with a spatial perception similar to that obtained with linear processing that served as the reference condition. The independent direct-sound driven compressor created a sense of movement of the sound between the two ears, suggesting that preserving the interaural level differences via linked compression is advantageous with the proposed direct-sound driven compression scheme.

[http://dx.doi.org/10.1121/1.4984040]

[GCS] Pages: 4556–4566

I. INTRODUCTION

In everyday acoustic environments, the sound that reaches a listener's ears contains the direct sound stemming from the different sound sources as well as reflections from obstacles in the surroundings. Despite the mixture of direct sound, early and late reflections that are typically present in rooms, normal-hearing listeners commonly perceive sound sources as being compact and correctly localized in space. It has been shown that both monaural cues, such as the sound pressure level at the ear drums and the direct-to-reverberant energy ratio (DRR) (Zahorik, 2002), as well as binaural cues, such as interaural time and level differences (Catic *et al.*, 2013; Hartmann and Wittenberg, 1996), contribute to reliable sound source localization in reverberant environments. Specifically, robust distance perception has been shown to be based on estimations of the DRR (Zahorik, 2005) whereas the sensation of externalized sound images, their azimuthal orientation in space and their apparent source width have been argued to be driven by binaural cues (e.g., Catic *et al.*, 2015; Whitmer *et al.*, 2012).

People with a sensorineural hearing impairment typically suffer from loudness recruitment, such that low-level sounds are not detectable while high-level sounds produce a close-to-normal loudness perception (e.g., Fowler, 1936; Steinberg and Gardner, 1937). To compensate for this reduced dynamic range of levels in the hearing-impaired listeners, level-dependent amplification is commonly applied in hearing aids, such that low-level sounds are amplified more than higher-level sounds (Allen, 1996). This corresponds to a compressive processing of the input level range to the smaller dynamic range of levels that can be perceived by the listener. If such dynamic range compression in hearing aids operates independently in the left-ear and right-ear channels, less amplification is typically provided to the ear signal that is closer to a given sound source than to the ear signal that is farther away from the sound source, such that the intrinsic interaural level differences (ILDs) in the sound are reduced. In anechoic conditions, this can lead to perceived lateral movements of the sound image (Wiggins and Seeber, 2011, 2012). To avoid this, state-of-the-art bilaterally fitted hearing aids share the measured sound intensity information across both devices via a wireless link (Korhonen *et al.*, 2015). This shared processing is commonly referred to as "linked" compression, such that in the case of a symmetrical hearing loss the amplification provided by the two compressors is the same in both ears and, as a consequence, the intrinsic ILDs are preserved. This has been shown to improve the ability of normal-hearing listeners to attend to a desired target in an auditory scene with spatially

[a)]Electronic mail: tdau@elektro.dtu.dk separated maskers as compared to independent compression in reverberant conditions (Schwartz and Shinn-Cunningham, 2013).

However, as demonstrated in Hassager *et al.* (2017) both independent and linked fast-acting compression (with an attack and release time of 10 and 60 ms, respectively) can strongly distort the spatial perception of sounds in reverberant acoustic environments. Both compression strategies can lead to an increased diffusiveness of the perceived sound and broader, sometimes internalized ("in the head"), sound images as well as sound-image splits. Such spatial distortions were observed both in normal-hearing and hearing-impaired listeners when either linked or independent compression was applied to the signals. It was demonstrated that the observed spatial distortions mainly resulted from the applied compression enhancing the level of the reflected sound relative to the level of the direct sound. It was concluded that compressive hearing-aid processing needs to maintain the energy ratio of the direct sound to the reflected sound in order to preserve the natural spatial cues in the acoustic scene.

Ideally, a dereverberation of the binaural room impulse responses (BRIRs) for each of the sound sources would be required to apply compression to the individual "dry" sound sources, followed by a convolution of the individual sound sources with the respective BRIRs to reintroduce and preserve the spatial chararcteristics of a given scene. It was shown by Hassager *et al.* (2017) that this approach provided the listener with an undistorted spatial perception. However, such idealized processing requires *a priori* knowledge of the dry source signals and the respective BRIRs, which limits the potential applicability of this type of processing to actual hearing-aid applications.

An alternative approach to preserving the natural spatial properties of a sound scene would be to effectively "linearize" the compressive processing by using time constants that are longer than the reverberation time. However, such processing would compromise the restoration of loudness perception obtainable by fast-acting compression (Strelcyk *et al.*, 2012). In the present study, it was investigated whether fast-acting compression that preserves the listener's spatial impression could be achieved by adaptively adjusting the time constant of the compressor depending on a binary decision reflecting direct-sound activity. The idea was to maintain fast-acting compression in time-frequency (T-F) units dominated by the direct sound while linearizing the processing via longer time constants of the compressor in T-F units dominated by reverberation.

If BRIR information was available, the short-term estimate of the signal-to-reverberant energy ratio (SRR) could be used to identify T-F units that are dominated by the direct sound. Specifically, the BRIR could be split into its direct and reverberant parts (Zahorik, 2002). Then, the energy ratio of the direct sound (the source signal convolved with the direct part of the BRIR) to the reverberant sound (the source signal convolved with the reverberant part of the BRIR) could be used as a decision metric. For a given criterion (e.g., SRR > 0 dB), an *a priori* classification could be performed to identify those T-F units that are dominated by the direct sound. However, this technique is not feasible in practical applications because the BRIRs are typically not available. Therefore, several "blind" algorithms have been developed to estimate the presence of reverberation in signals without *a priori* knowledge of the BRIRs. For example, the interaural coherence (IC) can be used to estimate the amount of reverberation in a signal since reverberation reduces the IC (e.g.; Thiergart *et al.*, 2012; Westermann *et al.*, 2013; Zheng *et al.*, 2015). Hazrati *et al.* (2013) developed an algorithm operating on monaural signals to identify direct-sound dominated T-F units by extracting a variance-based feature from the reverberant signal and comparing it to an adaptive threshold. The algorithm generates a binary T-F classification that was applied to the signal to suppress reverberation. The authors reported significant speech intelligibility improvements in cochlear-implant users.

The present study focused on the spatial perception of speech presented in an everyday reverberant environment. The speech signals were processed by fast-acting hearing-aid compression with and without a binary classification stage to linearize the processing of T-F units dominated by reverberation. Besides the classification using the short-term SRR based on *a priori* knowledge of the BRIRs, the blind classification method by Hazrati *et al.* (2013) was tested both in independent and linked compression settings of the simulated hearing aid. The compression without the binary classification stage corresponded to conventional compression schemes described in the literature (e.g., Kates, 2008), whereas the compression with the binary classification stage represented the proposed direct-sound driven compression system. Linear processing, i.e., level-independent amplification, was used as the reference condition. Only normal-hearing listeners participated in the present study. The main goal was to evaluate the feasibility of the approach motivated by the results from Hassager *et al.* (2017). To quantify the distortion of the spatial perception in the different conditions, the IC of the ear signals was used as an objective metric.

II. COMPRESSION SYSTEM

A. Algorithm overview

Figure 1 shows the block diagram of the proposed algorithm. Both the independent and linked hearing-aid compression systems were based on short-time Fourier transformations (STFTs) and operated in seven octave-spaced frequency channels. In the STFT block, the left- and right-ear signals, sampled at a rate of 48 000 Hz, were divided into overlapping frames of 512 samples (corresponding to ~10.7 ms) with a shift of 128 samples. Each frame was Hanning-windowed and zero padded to a length of 1024 samples and transformed into the frequency domain by applying a 1024-point discrete Fourier transform (DFT). In the left and right filterbank (FB), the power of the DFT bins was integrated into seven octave-wide frequency bands with center frequencies ranging from 125 Hz to 8 kHz. Similarly, the direct-sound classification stages (see Sec. II B) consisted of seven octave-wide frequency bands. The power and the corresponding binary classification of the seven frequency bands were used to estimate the gain level (see Sec. II C).

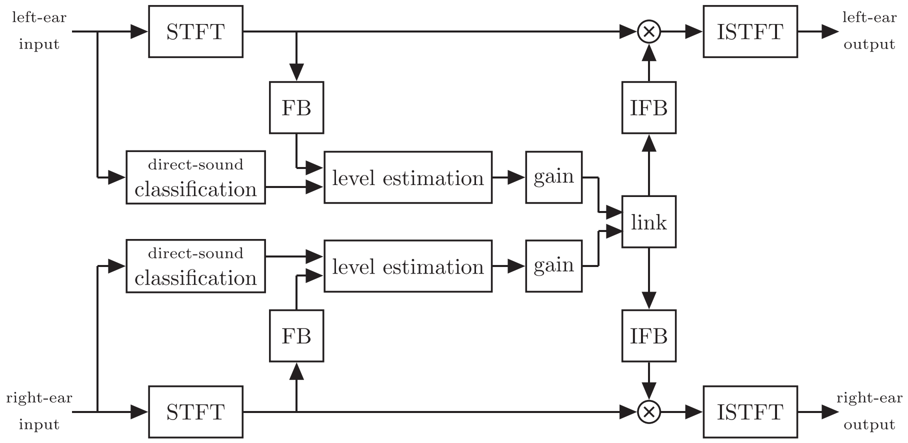

FIG. 1. Block diagram of the proposed direct-sound driven compressor. First the left- and right-ear signals are windowed in time segments and transformed into the frequency domain by a short-time Fourier transforms (STFT). The frequency bins in each time window are combined into seven octave spaced frequency bands by the filterbank (FB), thereby creating T-F units. In the direct sound classification block a binary classification is performed whether T-F units are dominated by the direct sound. In the level estimation and gain blocks, the T-F units are smoothed across time with time constants determined by the classification and the gain values for T-F units are found. In the link block, the gain values are either kept untouched or the minima of the left and right gain values are used as the gain values in both ears. In the inverse filterbank (IFB), the gains were then interpolated in the frequency domain and applied to the STFT bins of the input stimulus. Finally, an inverse STFT (ISTFT) was computed and the resulting temporal waveform was presented to the left and right ear.

The estimated levels for the individual T-F units were converted to sound pressure level (SPL) in dB, and a broken-stick gain function (with a linear gain below the compression threshold and a constant compression ratio above the threshold) was applied. The compression thresholds and compression ratios were calculated from NAL-NL2 prescription targets (Keidser *et al.*, 2011) for the N_3 audiogram corresponding to a flat and moderately sloping hearing-loss as defined in Bisgaard *et al.* (2010). The compression thresholds (CTs) and compression ratios (CRs) for the seven respective frequency bands are summarized in Table I.

The simulated input level to the compressor operating closest to the sound source was 75 dB SPL. In the case of independent processing, the gain values for the individual T-F units were kept untouched. In the case of linked processing, the minima of the left and right gain values were taken as the gain values in both ears. In the inverse filterbank (IFB), the resulting gains were then interpolated in the frequency domain using a piecewise cubic interpolation to avoid aliasing artifacts and applied to the STFT bins of the input stimulus. Finally, an inverse DFT of the STFT coefficients was computed to produce time segments of the compressed stimuli. These time segments were subsequently windowed with a tapered cosine window to avoid aliasing artifacts, and combined using an overlap-add method to provide the processed temporal waveform presented to the left and right ears.

Figure 2 illustrates the different processing stages of the proposed system in relation to a conventional compression system. Panel (a) shows anechoic speech at the output of an octave-wide bandpass filter tuned to 1000 Hz. Panel (b) shows the corresponding output for reverberant speech, illustrating the impact of reverberation on the dry source signal. The blind classification of direct-sound signal components is shown in panel (c) together with a conventional compressor using a fixed compression mode with short time constants (fast-acting). The gain functions of the proposed direct-sound driven compressor and the conventional compressor are shown in panel (d). Panel (e) shows the waveform of the compressed reverberant speech using the proposed direct-sound driven compressor, and panel (f) shows the waveform of the compressed reverberant speech processed with the conventional compressor. It is apparent that the conventional compressor amplifies the low-level portions of the sound and thereby enhances the reverberant components. In contrast, the proposed direct-sound driven compressor applies fast-acting compression in T-F units that are dominated by direct-sound components and slow-acting compression in T-F units that are dominated by reverberation.

TABLE I. The compression thresholds (CT) and compression ratios (CR) in the seven octave frequency bands.

	125 Hz	250 Hz	500 Hz	1000 Hz	2000 Hz	4000 Hz	8000 Hz
CT (dB SPL)	45	50	49	40	48	44	32
CR	3.4:1	3.2:1	2.3:1	2.7:1	3.6:1	3.8:1	4.0:1

B. Classification

The proposed direct-sound driven compressor requires a binary classification of individual T-F units into direct-sound and reverberant signal components. This classification was either based on the short-term SSR using *a priori* knowledge of the BRIRs or on the blind classification method described by Hazrati *et al.* (2013). The details of the two approaches are described below.

1. Signal-to-reverberant ratio classification

Assuming *a priori* knowledge about the BRIR, the short-term SRR was used as a decision metric to identify T-F units that are dominated by the direct sound. Specifically, the BRIRs were split into their direct and reverberant parts (Zahorik, 2002). The direct part was defined as the first 2.5 ms of the impulse response and the reverberant part was

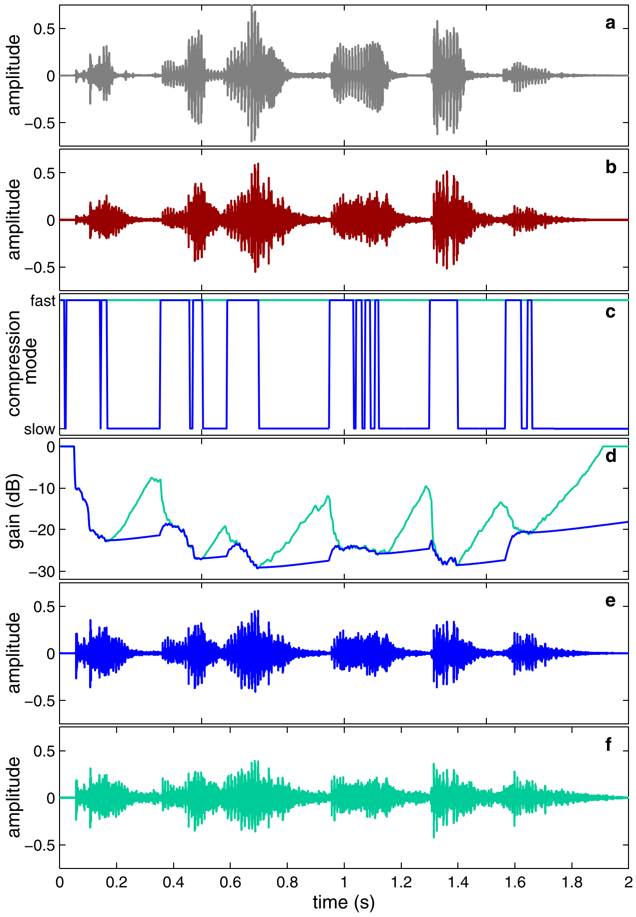

FIG. 2. (Color online) Example illustrating a bandpass filtered HINT sentence extracted at the center frequency of 1000 Hz. (a) Anechoic sentence, (b) reverberant sentence, (c) the blind binary classification (blue) where a value of one indicates direct-sound activity, (d) the corresponding gain function for conventional compression (light green) and the direct-sound driven compression (blue), (e) the reverberant sentence processed by the proposed direct-sound driven compression, and (f) the reverberant sentence processed by conventional compression.

defined as the remaining subsequent samples of the BRIRs. The 2.5 ms transition point was chosen here since the first reflection occurred immediately after this point in time. The reverberant part contained both the early reflections and the late reverberation. The direct signal and the reverberant signal were obtained by convolving the dry speech (source signal) with the direct part and the reverberant part of the BRIR, respectively. The direct signal, D, and the reverberant signal, R, were segmented into overlapping frames and decomposed into seven octave-wide frequency channels using the same parameters as the compressor. The power was thereafter smoothed in time (t) by recursive averaging as follows:

$$D_s(t,f) = \lambda D_s(t-1,f) + (1-\lambda)|D(t,f)|^2$$

and

$$R_s(t,f) = \lambda R_s(t-1,f) + (1-\lambda)|R(t,f)|^2,$$

where D_s and R_s represent the smoothed versions, and λ represents the smoothing constant which was determined by $\lambda = \exp(-k_{step}/(f_s\tau))$ for a time constant, τ, of 10 ms and a step size k_{step} of 128 samples at a sampling frequency f_s of 48 000 Hz. The SRR was calculated as

$$\mathrm{SRR}(t,f) = 10\log_{10}\left(\frac{D_s(t,f)}{R_s(t,f)}\right).$$

The classification of T-F units was performed by applying a local criterion to the short-term SSR, such that T-F units greater than 0 dB were assigned a value of one and zero otherwise, creating a binary SRR classification

$$C^{SRR}(t,f) = \begin{cases} 1, & SRR(t,f) > 0, \\ 0, & \text{otherwise}. \end{cases}$$

2. Blind classification

The blind detection of direct-sound components without prior knowledge was performed using the method described by Hazrati *et al.* (2013). The reverberant signal was band-pass filtered by seven octave-spaced filters to match the frequency resolution of the compressor. The band-pass filtered signals were then segmented into overlapping frames, denoted by S, and a variance-based feature labeled as F was calculated. The feature was computed by calculating the variance of the signal raised to a power, α, and dividing it by the variance of the absolute value of the signal. This ratio was then converted to dB:

$$F(t,f) = 10\log_{10}\left(\frac{\sigma^2(|S(t,f)|^\alpha)}{\sigma^2(|S(t,f)|)}\right),$$

where the exponent, α, was set to 1.75. This variance-based feature was then smoothed across time using a three-point median filter.

To obtain the binary classification of speech activity, C^{Blind}, the variance-based feature, F, was compared to an adaptive threshold T:

$$C^{Blind}(t,f) = \begin{cases} 1, & F(t,f) > T, \\ 0, & \text{otherwise}. \end{cases}$$

The adaptive threshold was based on the nonparametric and unsupervised method described by Otsu (1979) and ensured a robust classification in a variety of acoustic conditions. The adaptive threshold was calculated for each T-F unit separately and involved a histogram analysis of the variance-based feature across a temporal context of 80 ms.

3. Classification parameters

The parameters of the blind classification, including the exponent, α, and the temporal context exploited by the adaptive threshold, were adjusted to account for an SRR threshold criterion of 0 dB, as opposed to a local criterion of −8 dB that was used in the study by Hazrati *et al.* (2013). To quantify the performance of the blind classification, the hit rate minus the false-alarm rate (H-FA) was computed by comparing the detection of direct-sound components to the short-term SRR classification in the seven frequency channels. Clean training sentences from the Danish hearing in noise test corpus (Danish HINT) (Nielsen and Dau, 2011) were randomly selected and convolved with BRIRs corresponding to room A and room B from the Surrey database (Hummersone *et al.*, 2010). The Surrey database was recorded with a Cortex head and torso simulator (HATS). Room A ($T_{60} = 0.32$ s and $DRR = 6.09$ dB) and room B ($T_{60} = 0.49$ s and $DRR = 5.31$ dB) represent acoustic environments with moderate reverberation. However, as described in Sec. III B, the direct-sound driven compressor was tested in an IEC listening room with individual HRTFs, requiring that the blind classification approach generalizes to unseen HRTFs and unseen room conditions. The evaluation was performed using all 37 azimuth angles ranging from −90° to 90°. The results were averaged across rooms and azimuth angles and are shown in Table II. The hit rate (H) was defined as the percentage of correctly classified direct-sound dominant T-F units, while the false-alarm rate (FA) was defined as the percentage of wrongly classified T-F units dominated by reverberation. Apart from the two lowest frequency bands (at 125 Hz and 250 Hz), where the FAs are higher than at all other frequencies, the blind classification produced a reasonably high performance in terms of the H-FA metric, given that the chance for H-FA is 0%.

C. Level estimation

The levels of the T-F units were estimated by smoothing the power of the T-F units across time using recursive averaging:

$$X_s(t,f) = cX_s(t-1,f) + (1-c)|X(t,f)|^2,$$

where $|X|^2$ represents the power of the individual T-F units, X_s the smoothed power, and c the smoothing constant. The smoothing constant, c, was updated according to the following criteria:

$$c = \begin{cases} c^{fast}_{attack}, & \text{when} |X(t,f)|^2 \geq X_s(t-1,f) \text{ and } \mathrm{C}(t,f)=1, \\ c^{fast}_{release}, & \text{when} |X(t,f)|^2 < X_s(t-1,f) \text{ and } \mathrm{C}(t,f)=1, \\ c^{slow}_{attack}, & \text{when} |X(t,f)|^2 \geq X_s(t-1,f) \text{ and } \mathrm{C}(t,f)=0, \\ c^{slow}_{release}, & \text{when} |X(t,f)|^2 < X_s(t-1,f) \text{ and } \mathrm{C}(t,f)=0, \end{cases}$$

with C either $\mathrm{C}^{\mathrm{SRR}}$ or $\mathrm{C}^{\mathrm{Blind}}$ and the smoothing constants, c^{fast}_{attack}, $c^{fast}_{release}$, c^{slow}_{attack}, and $c^{slow}_{release}$, found according to IEC 60118-2 (1983), to be 10, 60, 2000, and 2000 ms, respectively. When C is equal to one the compression mode is fast-acting and when C is equal to zero the compression mode is slow-acting.

TABLE II. The blind classification performance in terms of the H, HA, and H-FA for the seven octave frequency channels averaged across rooms and azimuth angles.

Frequency	125 Hz	250 Hz	500 Hz	1000 Hz	2000 Hz	4000 Hz	8000 Hz
H	95.5%	99.3%	98.5%	96.8%	92.5%	78.7%	87.1%
FA	57.7%	54.2%	40.9%	36.0%	28.7%	11.8%	26.3%
H-FA	38.4%	44.8%	57.6%	60.9%	64.0%	66.9%	60.7%

III. METHODS

A. Listeners

Eighteen normal-hearing listeners (10 males and 8 females), aged between 19 and 35 years, participated in the experiment. All had audiometric pure-tone thresholds below 20 dB hearing level at frequencies between 125 Hz and 8 kHz. All listeners signed an informed consent document and were reimbursed for their efforts.

B. Experimental setup and procedure

The experimental setup and procedure were similar to the ones described in Hassager *et al.* (2017). The experiments took place in a reverberant listening room designed in accordance with the IEC 268-13 (1985) standard. The room had a reverberation time T_{30} of approximately 500 ms, corresponding to a typical living room environment. Figure 3 shows the top view of the listening room and the experimental setup as placed in the room. The dimensions of the room were 752 cm × 474 cm × 276 cm (L × W × H). Twelve Dynaudio BM6 loudspeakers were placed in a circular arrangement with a radius of 150 cm, distributed with equal spacing of 30° on the circle. A chair with a headrest and a Dell s2240t touch screen in front of it were placed in the center of the loudspeaker ring. The listeners were seated on the chair with view direction to the loudspeaker placed at 0° azimuth. The chair was positioned at a distance of 400 cm from the wall on the left and 230 cm from the wall behind.

The graphical representation of the room and the setup, as illustrated in Fig. 3, were also shown on the touch screen, without the information regarding the room dimensions. In addition to the loudspeakers, a Fireface UCX sound card operating at a sampling frequency of 48 000 Hz, two DPA high sensitivity microphones and a pair of HD850 Sennheiser headphones were used to record the individual BRIRs for the listeners (see Sec. III C). The BRIRs were measured from the loudspeakers placed at the azimuth angles of 0° and 300°. The listeners were instructed to support the back of their head on the headrest while remaining still and to fixate on a

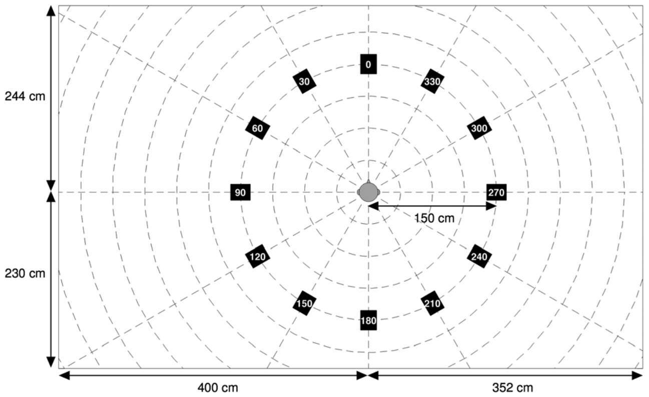

FIG. 3. The top view of the experimental setup. The loudspeaker positions are indicated by the black squares. The grey circle in the center indicates the position of the chair, where the listener was seated. The listeners had a view direction on the loudspeaker placed at the 0° degree azimuth. The graphical representation was also shown on the touch screen, without the room dimensions shown in the figure.

marking located straight ahead (0°) both during the BRIR measurements and during the sound presentations. On the touch screen, the listeners were asked to place circles on the graphical representation as an indication of the perceived position and width of the sound image in the horizontal plane. By placing a finger on the touch screen, a small circle appeared on the screen with its center at the position of the finger. When moving the finger while still touching the screen, the circumference of the circle would follow the finger. When the desired size of the circle was reached, the finger was released from the screen. By touching the center of the circle and moving the finger while touching the screen, the position of the circle would follow along. By touching the circumference of the circle and moving the finger closer to or farther away from the center of the circle while touching the screen, the circle would decrease or increase in size, respectively. A double tap on the center of the circle would delete the circle. If the listeners perceived a split of any parts of the sound image, they were asked to place multiple circles reflecting the positions and widths of the split images. The listeners were instructed to ignore other perceptual attributes, such as sound coloration and loudness. Each stimulus was presented three times from each of the two loudspeaker positions. No response feedback was provided to the listeners. The test conditions and active loudspeaker position were presented in random order within each run.

C. Spatialization

Individual BRIRs were measured to simulate the different conditions virtually over headphones. Individual BRIRs were used because it has been shown that the use of individual head-related transfer functions (HRTFs), the Fourier transformed head-related impulse responses, improve sound localization performance compared to non-individual HRTFs (e.g., Majdak *et al.*, 2014), as a result of substantial cross-frequency differences between the individual listeners' HRTFs (Middlebrooks, 1999). Individual BRIRs were measured from the loudspeakers placed at the azimuth angles of 0° and 300°. The BRIR measurements were performed as described in Hassager *et al.* (2017). The microphones were placed at the ear-canal entrances and were securely attached with strips of medical tape. A maximum-length-sequence (MLS) of order 13, with 32 repetitions played individually from each of the loudspeakers, was used to obtain the impulse response, h_{brir}, representing the BRIR for the given loudspeaker. The headphones were placed on the listeners and corresponding headphone impulse responses, h_{hpir}, were obtained by playing the same MLS from the headphones. To compensate for the headphone coloration, the inverse impulse response, h_{hpir}^{inv}, was calculated in the time domain using the Moore-Penrose pseudoinverse. By convolving the room impulse responses, h_{brir}, with the inverse headphone impulse responses, h_{hpir}^{inv}, virtualization filters with the impulse responses, h_{virt}, were created. Stimuli convolved with h_{virt} and presented over the headphones produced the same auditory sensation in the ear-canal entrance as the stimuli presented by the loudspeaker from which the filter, h_{brir}, had been recorded. Hence, a compressor operating on an acoustic signal convolved with h_{brir} behaves as if it was implemented in a completely-in-canal hearing aid.

To validate the BRIRs, the stimuli were played first from the loudspeakers and then via the headphones filtered by the virtual filters h_{virt}. In this way, it could be tested whether the same percept was obtained when using loudspeakers or headphones. By visual inspection, the graphical responses obtained with the headphone presentations were compared to the graphical responses obtained with the corresponding loudspeaker presentations. This comparison confirmed that all listeners had a very similar spatial perception in the two conditions (see also Hassager *et al.*, 2017).

D. Stimuli and processing conditions

Speech sentences from the Danish HINT (Nielsen and Dau, 2011) were used as stimuli. The clean speech signals were convolved with the listener's BRIRs, h_{brir}, and then processed by the compression conditions. As listed in Table III, a set of six different compressor systems were tested: (1) Conventional independent compression that processed the binaural signals independently, (2) conventional linked compression that synchronizes the processing of the binaural signals, (3) independent compression with an SSR classification stage, (4) independent compression with a blind classification stage, (5) linked compression with an SSR classification stage, (6) linked compression with a blind classification stage. Linear processing was used as a reference condition. To compensate for the effect of the headphones, the left- and right-ear signals were afterwards convolved with the left and right parts of h_{hpir}^{inv}, respectively. The SPL of the stimulus at the ear closest to the sound source was 65 dB in all conditions.

E. Statistical analysis

The graphical responses provided a representation of the perceived sound image in the different conditions. To quantify deviations in the localization from the loudspeaker position across the different conditions, the root-mean-square (RMS) error of the Euclidean distance from the center of the circles to the loudspeakers was calculated. To reduce the confounding influence of front-back confusions as a result of the virtualization method, the responses placed in the opposite hemisphere (front versus rear) of the virtually playing loudspeaker were reflected across the interaural axis to the mirror symmetric position.

TABLE III. Overview of the different processing conditions involving compression.

Method	Binaural link	Compression Mode	Estimator
Independent	Off	Conventional	—
Linked	On	Conventional	—
Independent SRR	Off	Direct-sound driven	Short-term SRR
Independent blind	Off	Direct-sound driven	Blind
Linked SRR	On	Direct-sound driven	Short-term SRR
Linked blind	On	Direct-sound driven	Blind

An analysis of variance (ANOVA) was conducted on two mixed-effect models to evaluate whether the processing condition and loudspeaker position had an effect on the dependent variable, which was either the RMS error or the radius of the placed circles. In the mixed-effect models, listeners were treated as a random block effect nested within the repeated within-listener measures of repetition, processing condition and loudspeaker position. Repetitions were treated as a random effect, while the processing condition and loudspeaker position were treated as fixed effects. The radius data were square-root transformed and the RMS error was log transformed to correct for heterogeneity of variance. The assumptions underlying parametric analysis was met after the transformations. Tukey's HSD corrected *post hoc* tests were conducted to test for main effects and interactions. A confidence level of 5% was considered to be statistically significant, and only statistically significant results are reported.

F. Analysis of spatial cues

In order to quantify the effect of the different compression schemes on the spatial cues, the interaural coherence (IC) was calculated. The IC can be defined as the absolute maximum value of the normalized cross-correlation between the left- and right-ear output signals $s_{out,l}$ and $s_{out,r}$ occurring over an interval of $|\tau| \leq 1$ ms (e.g., Blauert and Lindemann, 1986; Hartmann *et al.*, 2005):

$$\mathrm{IC} = \max_{\tau} \left| \frac{\sum_t s_{out,l}(t+\tau)\, s_{out,r}(t)}{\sqrt{\sum_t s_{out,l}^2(t) \sum_t s_{out,r}^2(t)}} \right|.$$

For each individual listener, the left- and right-ear output signals were filtered with an auditory inspired "peripheral" filterbank consisting of complex fourth-order gammatone filters with equivalent rectangular bandwidth spacing (Glasberg and Moore, 1990). The IC was subsequently computed from the filtered output signals. The just-noticeable difference (JND) in IC is about 0.04 for an IC equal to 1 and increases to 0.4 for an IC equal to 0 (Gabriel and Colburn, 1981; Pollack and Trittipoe, 1959). The IC distribution was estimated by applying a Gaussian kernel-smoothing window with a width of 0.02 (half of the smallest JND) to the IC histograms.

IV. RESULTS

A. Experimental data

Figures 4 and 5 show graphical representations of the listeners' responses, including repetitions, virtualized from the loudspeaker positioned at 300° azimuth. The pattern of results obtained at the loudspeaker positioned at 0° azimuth was similar to that observed for the loudspeaker positioned at 300°. The data for 0° are provided in the supplementary material.[1] In Fig. 4, the upper left panel represents the responses for the linear processing (reference) condition,

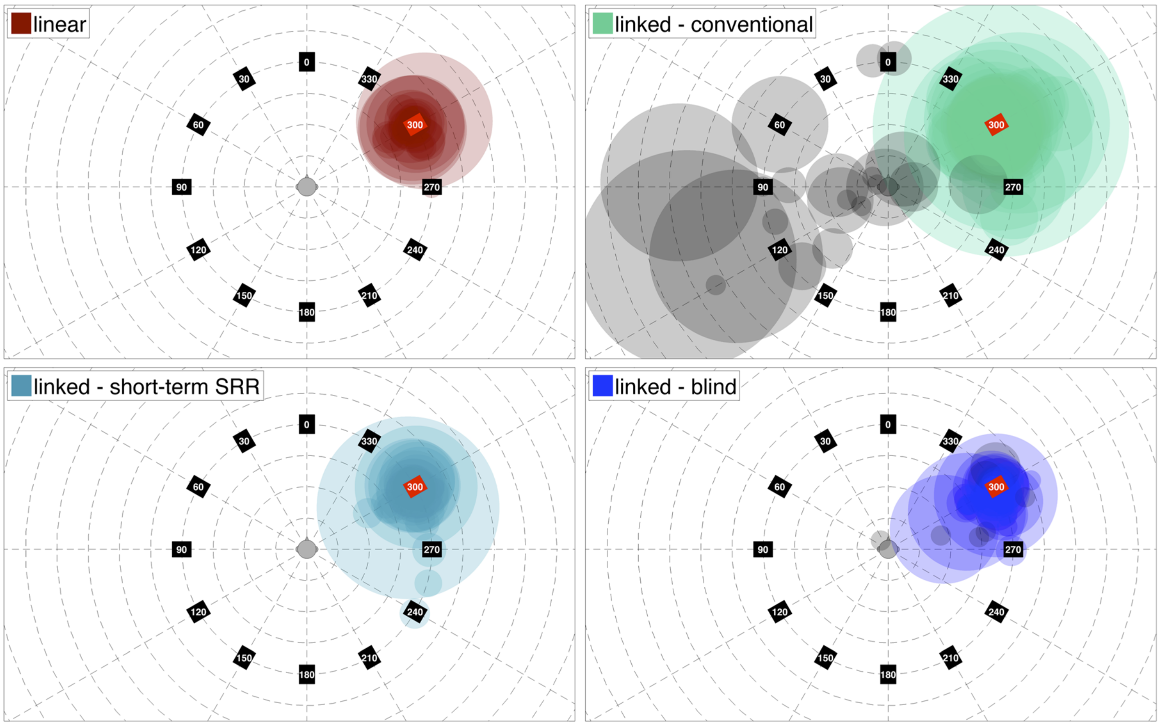

FIG. 4. (Color online) Graphical representations of the listeners' responses obtained with the speech virtually presented from the 300° position in the listening room. The upper left panel shows the results for linear processing (reference condition). The results for conventional linked compression, direct-sound driven linked compression based on SRR classification, and direct-sound driven linked compression based on blind classification are shown in the upper right, lower left, and lower right panels, respectively. The response of each individual listener is indicated as a transparent filled circle with a center and width corresponding to the associated perceived sound image. The main sound images are indicated by the different colors in the different conditions whereas split images are indicated in gray.

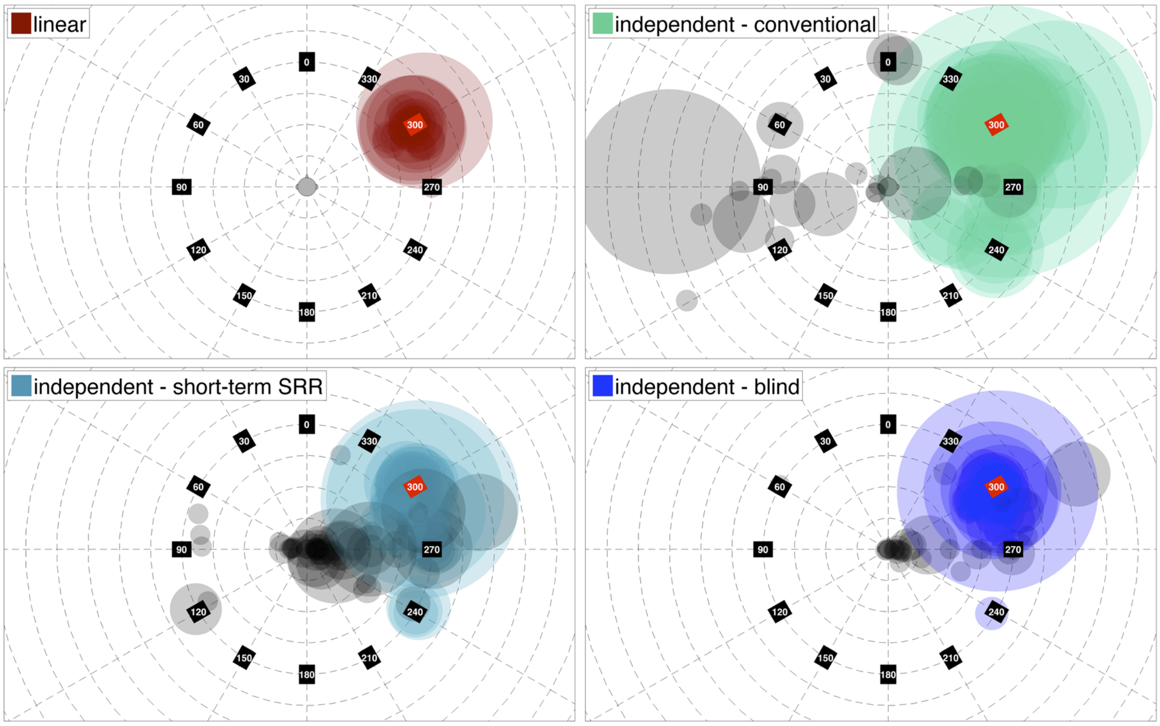

FIG. 5. (Color online) Same as Fig. 4, but for the independent compression conditions.

whereas the responses obtained with conventional linked compression, direct-sound driven linked compression based on SRR classification and direct-sound driven linked compression based on blind classification are shown in the upper right, lower left and lower right panel, respectively. The responses of each individual listener in a given condition are indicated as transparent filled (colored and gray) circles with a center and size corresponding to the associated perceived sound image in the top-view perspective of the listening room (including the loudspeaker ring and the listening position in the center of the loudspeakers). Overlapping areas of circles obtained from different listeners are reflected by the increased cumulative intensity of the respective color code. To illustrate when a listener experienced a split in the sound image and, therefore, indicated more than one circle on the touch screen, only the circle the listener placed nearest to the loudspeaker (including positions obtained by front-back confusions) was indicated in color whereas the remaining locations were indicated in gray.

In the reference condition (upper left panel in Fig. 4), the sound was perceived as coming from the loudspeaker position at 300° azimuth. In contrast, in the conventional linked compression condition (upper right panel), the sound was generally perceived as being wider and, in some cases, as occurring closer to the listener than the loudspeaker or between the loudspeakers at 240° and 300° azimuth. For some of the listeners, the conventional linked compression also led to split images as indicated by the gray circles. These results are consistent with the results obtained in Hassager *et al.* (2017). In the direct-sound driven linked compression conditions based on SRR classification (lower left panel) and blind classification (lower right panel), the listeners perceived the sound image as being compact and located mainly at the loudspeaker at 300° azimuth. None of the listeners experienced image splits with the direct-sound driven compression based on the SRR classification, while some image splits were experienced with the direct-sound driven compression using the blind classification. Nonetheless, in contrast to the conventional linked compression, the experienced image splits were concentrated mainly in the region around the loudspeaker that the sound was virtualized from.

Figure 5 shows the corresponding results for independent compression. The general pattern of results was similar to that found for linked compression (from Fig. 4). However, the responses for direct-sound driven independent compression based on the SRR classification (lower left panel) and the blind classification (lower right panel) contained considerably more image splits than the corresponding responses for conventional linked compression (upper right panel of Fig. 4). The reported image splits were in both direct-sound driven compression conditions placed around the position of the head. The listeners who indicated image splits reported verbally that they perceived an internalized sense of movement of the sound between the two ears. Nonetheless, the listeners generally perceived the main sound as being compact and located mainly at the loudspeaker at 300° azimuth in the two classification conditions.

For the radius of the placed circles, indicating the perceived width of the sound image, the ANOVA revealed an effect of processing condition $[F(6, 42) = 65.62, p \ll 0.001]$ and an interaction between processing condition and loudspeaker position $[F(6, 607) = 3.86, p < 0.001]$. *Post hoc* comparisons revealed significant differences between conventional compression and direct-sound driven compression $[p \ll 0.001]$, and between conventional compression and linear processing $[p \ll 0.001]$. This was found for the linked as well as the independent condition. The mean radii in the conventional compression conditions were 34.6 and 37.0 cm for the linked and the independent compression condition, respectively, while the mean radii in the other conditions were between 3.3 and 9.1 cm. Significantly higher radius (1 cm) was found for the 300° azimuth loudspeaker position than for the frontal loudspeaker position for linked direct-sound driven compression. No other significant differences in radius were found between the loudspeaker positions for the other processing conditions. For the RMS error, the ANOVA showed an effect of the loudspeaker position $[F(1, 17) = 6.82, p = 0.02]$. *Post hoc* comparisons showed that the RMS error was slightly higher at the 300° azimuth loudspeaker position than at the frontal loudspeaker position. This is consistent with previous studies (e.g., Mills, 1958) demonstrating a higher localization acuity for frontal than for lateral positioned sound sources.

B. Analysis of spatial cues

Figure 6 shows the IC distributions for linear processing and the linked compression conditions (conventional, direct-sound driven with either SRR or blind classification) for the speech virtualized from the frontal loudspeaker. For simplicity, only the results at the output of the gammatone filter tuned to 1000 Hz are shown. The IC distributions for the linear processing (solid red line) and the direct-sound driven linked compression with either short-term SRR (dashed light blue line) or blind classification (dashed blue line) are similar to each other whereas the distribution for the conventional linked compression (dashed light green line) has its maximum at a much lower value. The distribution obtained with the linear processing shows a maximum at an IC of about 0.85. In contrast, the maxima of the distributions for the conventional linked compression condition are shifted towards a lower value of about 0.79. The same trends were observed for the independent compression conditions (not shown explicitly).

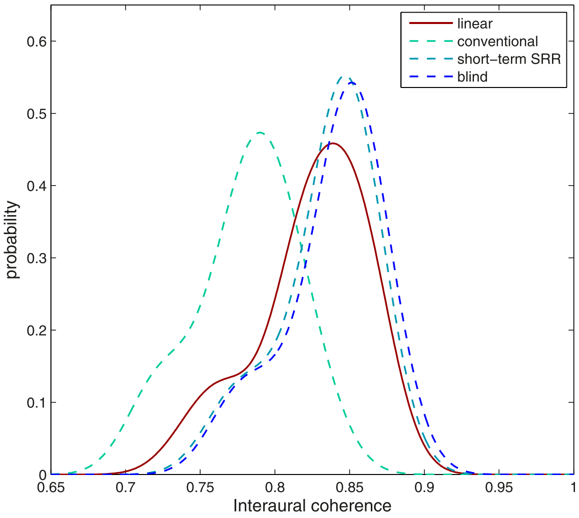

FIG. 6. (Color online) IC distributions of the ears signals, pooled across all listeners, at the output of the gammatone filter tuned to 1000 Hz. Results are shown for the speech virtualized from the frontal loudspeaker position. The solid red, dashed light green, dashed light blue and dashed blue curves represent the IC distributions for linear processing, conventional linked compression, direct-sound driven linked compression with SRR classification, and direct-sound driven linked compression with blind classification, respectively.

V. DISCUSSION

The present study compared conventional (independent and linked) fast-acting compression with direct-sound driven (independent and linked) compression. The classification stage in the direct-sound driven compressor was either based on the short-term SRR using *a priori* knowledge of the BRIRs or on the blind classification method by Hazrati *et al.* (2013). A spatial cue analysis showed that, in an everyday reverberant environment, conventional compression markedly reduced the IC of the stimulus between the ears relative to linear processing. The reason for this reduction is that the segments of the stimuli dominated by reverberation often exhibit a lower signal level and are therefore amplified stronger by the compression scheme than the stimulus segments that are dominated by the direct sound [see also Hassager *et al.* (2017)]. In contrast, the IC was largely maintained in the case of the direct-sound driven compression schemes relative to linear processing, implying that the energy ratio of the direct-sound to reverberation was preserved by linearizing the processing of the T-F units that are dominated by reverberation.

Consistent with the IC analysis, the direct-sound driven linked compression provided the listeners with a similar spatial percept as the linear processing scheme, while the conventional linked compression resulted in more diffuse and broader sound images as well as image splits. In the independent compression conditions, the general pattern of results was similar to that found for linked compression, except that the direct-sound driven compressor in the independent configuration led to the perception of an internalized sound image that is moving between the two ears. Previous studies have demonstrated that, in anechoic conditions, independent compression can lead to such perceived lateral movements of the sound image (Wiggins and Seeber, 2011, 2012), probably due to slow ILD changes over time. Interestingly, according to the verbal reports of most of the listeners in the present study, the sense of movement was not experienced in the case of the conventional independent compression condition, potentially because in this condition the increased amount of reverberation masks the occurrence of the ILD distortions stemming from the direct sound.

Instead of reconstructing the anechoic source signal, which would allow for the application of a “spatially ideal” compressor (Hassager *et al.*, 2017), the proposed compression scheme utilizes short-term estimates of direct-sound components as a control signal to adaptively select the appropriate time constants, thus avoiding artifacts and signal distortions inevitably introduced by dereverberation algorithms. The results indicated that the proposed processing scheme does not introduce artifacts other than the enhanced reverberation due to misclassification of reverberant components. The performance analysis of the blind classification revealed that fast-acting compression, in fact, is applied to T-F units dominated by the direct sound, as reflected in the observed large hit rates, whereas the T-F units dominated by reverberation are classified less accurately, as represented by the false alarm rates (see Table II). Nevertheless, the behavioral results did not show significant spatial distortions in the two linked direct-sound driven compression schemes, indicating that the binary classification performance and thereby the ability of the blind classification approach to generalize to unseen acoustic environments was reasonably high.

The experiments were conducted on normal-hearing listeners who have normal loudness perception and thus do not need level-dependent amplification, i.e., hearing-aid compression. Normal-hearing listeners were considered here because Hassager *et al.* (2017) demonstrated that hearing-aid compression affected hearing-impaired and normal-hearing listeners to a similar degree. Whereas the hearing-impaired listeners showed generally less accurate localization ratings than the normal-hearing listeners, the distortions resulting from conventional compression dominated the results and were similar in both listener groups. However, it will of course be crucial to perform corresponding experiments with the proposed direct-sound driven compression system with hearing-impaired listeners to further evaluate its significance and effectiveness. Furthermore, in the experiments considered in the present study, only a single sound source was used. With several sound sources, the impact of distorted spatial cues by conventional compression may limit the benefit that users are able to gain from current hearing aids. Thus, studying the influence of the direct-sound driven compression in multi-source scenario will be highly relevant. The blind estimation might be able to provide a robust estimation of direct-sound activity in multi-source scenarios because it does not require knowledge about the number or the spatial distribution of the sound sources.

There are certainly various ways to improve the detection of direct-sound components, e.g., by combining the monaural cues employed by Hazrati's method with binaural cues, such as the interaural coherence. Moreover, the adaptive threshold could be replaced by supervised learning approaches which were shown to enable accurate sound source localization in multi-source environments (May *et al.*, 2011, 2015). The present study was not focused on providing an optimized "solution" and parameter set of a compression system. Instead, the main goal was to demonstrate the principal effect of a compression system that is controlled via the surrounding reverberation statistics, such that the spatial perception of the acoustic scene becomes less distorted by the effects of compression on the reverberant portions of the ears' input signals.

VI. CONCLUSION

This study presented a direct-sound driven compression scheme that applied fast-acting compression in T-F units dominated by the direct sound while linearizing the processing via longer time constants in T-F units dominated by reverberation. It was demonstrated that such a direct-sound driven compression scheme can strongly reduce spatial distortions that are introduced by conventional compressors due to the enhancement of reverberant energy. It was found that linked direct-sound driven compression provided the listeners with a spatial percept similar to that obtained with linear processing. This was confirmed by the interaural coherence of the ear signals that was similar to that in the case of linear processing. A blind classification method was shown to provide accurate classification of direct-sound dominated T-F units. The blind classification method's performance was similar to that obtained with a classification based on the short-term SRR using *a priori* knowledge of the BRIRs. In general, such a classification stage was found to be necessary and ensured that fast-acting compression was only applied to the speech signal. The T-F units dominated by reverberation were classified less accurately which, however, did not produce a detrimental effect on the spatial perception ratings. In addition, it was found that, in the conditions with independent direct-sound driven compression, a sense of movement of the sound between the two ears was observed. Thus, linking the left- and right-ear compression in combination with the proposed direct-sound driven compression scheme might be a successful strategy to provide a natural spatial perception while restoring loudness as perceived by normal-hearing listeners.

ACKNOWLEDGMENT

This project was supported by Widex, Oticon, GN ReSound, and the Technical University of Denmark in connection to the Centre for Applied Hearing Research (CAHR). The authors wish to thank the reviewers for their helpful and constructive feedback regarding this work.

[1] See supplementary material at http://dx.doi.org/10.1121/1.4984040 for the graphical representations of the listeners' responses, including repetitions, virtualized from the loudspeaker positioned at 0 degree azimuth.

Allen, J. B. (**1996**). "Derecruitment by multiband compression in hearing aids," in *Psychoacoustics, Speech, and Hearing Aids* (World Scientific, Singapore), p. 372.

Bisgaard, N., Vlaming, M. S. M. G., and Dahlquist, M. (**2010**). "Standard audiograms for the IEC 60118-15 measurement procedure," Trends Amplif. **14**, 113–120.

Blauert, J., and Lindemann, W. (**1986**). "Spatial mapping of intracranial auditory events for various degrees of interaural coherence," J. Acoust. Soc. Am. **79**, 806–813.

Catic, J., Santurette, S., Buchholz, J. M., Gran, F., and Dau, T. (**2013**). "The effect of interaural-level-difference fluctuations on the externalization of sound," J. Acoust. Soc. Am. **134**, 1232–1241.

Catic, J., Santurette, S., and Dau, T. (**2015**). "The role of reverberation-related binaural cues in the externalization of speech," J. Acoust. Soc. Am. **138**, 1154–1167.

Fowler, E. P. (**1936**). "A method for the early detection of otosclerosis: A study of sounds well above threshold," Arch. Otolaryngol. Head Neck Surg. **24**, 731–741.

Gabriel, K. J., and Colburn, S. H. (**1981**). "Interaural correlation discrimination: I. Bandwidth and level dependence," J. Acoust. Soc. Am. **69**, 1394–1401.

Glasberg, B. R., and Moore, B. C. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103–138.

Hartmann, W. M., Rakerd, B., and Koller, A. (**2005**). "Binaural coherence in rooms," Acta Acust. Acust. **91**, 451–462.

Hartmann, W. M., and Wittenberg, A. (**1996**). "On the externalization of sound images," J. Acoust. Soc. Am. **99**, 3678–3688.

Hassager, H. G., Wiinberg, A., and Dau, T. (**2017**). "Effects of hearing-aid dynamic range compression on spatial perception in a reverberant environment," J. Acoust. Soc. Am. **141**, 2556–2568.

Hazrati, O., Lee, J., and Loizou, P. C. (**2013**). "Blind binary masking for reverberation suppression in cochlear implants," J. Acoust. Soc. Am. **133**, 1607–1614.

Hummersone, C., Mason, R., and Brookes, T. (**2010**). "Dynamic precedence effect modeling for source separation in reverberant environments," IEEE Trans. Audio. Speech. Lang. Process. **18**, 1867–1871.

IEC 268-13 (**1985**). "Sound system equipment. Part 13: Listening tests on loudspeaker" (International Electrotechnical Commission, Geneva, Switzerland).

IEC 60118-2 (**1983**). "Hearing aids. Part 2: Hearing aids with automatic gain control circuits" (International Electrotechnical Commission, Geneva, Switzerland).

Kates, J. M. (**2008**). *Digital Hearing Aids* (Plural, San Diego, CA).

Keidser, G., Dillon, H. R., Flax, M., Ching, T., and Brewer, S. (**2011**). "The NAL-NL2 prescription procedure," Audiol. Res. **1**(e24), 88–90.

Korhonen, P., Lau, C., Kuk, F., Keenan, D., and Schumacher, J. (**2015**). "Effects of coordinated compression and pinna compensation features on horizontal localization performance in hearing aid users," J. Am. Acad. Audiol. **26**, 80–92.

Majdak, P., Baumgartner, R., and Laback, B. (**2014**). "Acoustic and non-acoustic factors in modeling listener-specific performance of sagittal-plane sound localization," Front. Psychol. **5**, 1–10.

May, T., Ma, N., and Brown, G. J. (**2015**). "Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues," in *IEEE International Conference on Acoustics and Speech Signal Processing (ICASSP)*, pp. 2679–2683.

May, T., van de Par, S., and Kohlrausch, A. (**2011**). "A probabilistic model for robust localization based on a binaural auditory front-end," IEEE Trans. Audio. Speech. Lang. Process. **19**, 1–13.

Middlebrooks, J. C. (**1999**). "Individual differences in external-ear transfer functions reduced by scaling in frequency," J. Acoust. Soc. Am. **106**, 1480–1492.

Mills, A. W. (**1958**). "On the minimum audible angle," J. Acoust. Soc. Am. **30**, 237–246.

Nielsen, J. B., and Dau, T. (**2011**). "The Danish hearing in noise test," Int. J. Audiol. **50**, 202–208.

Otsu, N. (**1979**). "A threshold selection method from gray-level histograms," IEEE Trans. Syst. Man. Cybernetics. **9**, 62–66.

Pollack, I., and Trittipoe, W. (**1959**). "Interaural noise correlations: Examination of variables," J. Acoust. Soc. Am. **31**, 1616–1618.

Schwartz, A. H., and Shinn-Cunningham, B. G. (**2013**). "Effects of dynamic range compression on spatial selective auditory attention in normal-hearing listeners," J. Acoust. Soc. Am. **133**, 2329–2339.

Steinberg, J., and Gardner, M. (**1937**). "The dependence of hearing impairment on sound intensity," J. Acoust. Soc. Am. **9**, 11–23.

Strelcyk, O., Nooraei, N., Kalluri, S., and Edwards, B. (**2012**). "Restoration of loudness summation and differential loudness growth in hearing-impaired listeners," J. Acoust. Soc. Am. **132**, 2557–2568.

Thiergart, O., Del Galdo, G., and Habets, E. A. P. (**2012**). "Signal-to-reverberant ratio estimation based on the complex spatial coherence between omnidirectional microphones," in *IEEE International Conference on Acoustics and Speech Signal Processing (ICASSP)*, pp. 309–312.

Westermann, A., Buchholz, J. M., and Dau, T. (**2013**). "Binaural dereverberation based on interaural coherence histograms," J. Acoust. Soc. Am. **133**, 2767–2777.

Whitmer, W. M., Seeber, B. U., and Akeroyd, M. A. (**2012**). "Apparent auditory source width insensitivity in older hearing-impaired individuals," J. Acoust. Soc. Am. **132**, 369–379.

Wiggins, I. M., and Seeber, B. U. (**2011**). "Dynamic-range compression affects the lateral position of sounds," J. Acoust. Soc. Am. **130**, 3939–3953.

Wiggins, I. M., and Seeber, B. U. (**2012**). "Effects of dynamic-range compression on the spatial attributes of sounds in normal-hearing listeners," Ear Hear. **33**, 399–410.

Zahorik, P. (**2002**). "Direct-to-reverberant energy ratio sensitivity," J. Acoust. Soc. Am. **112**, 2110–2117.

Zahorik, P. (**2005**). "Auditory distance perception in humans: A summary of past and present research," Acta Acust. Acust. **91**, 409–420.

Zheng, C., Schwarz, A., Kellermann, W., and Li, X. (**2015**). "Binaural coherent-to-diffuse-ratio estimation for dereverberation using an ITD model," in *European Signal Processing Conference (EUSIPCO)*, pp. 1048–1052.