A hear-through system for plausible auditory contrast enhancement

Marian Weger
Robert Höldrich
Institute for Electronic Music and Acoustics (IEM)
University of Music and Performing Arts
Graz, Austria
weger@iem.at

ABSTRACT

In many of our everyday and professional routines, we rely on knowledge we gather from the auditory feedback of physical interactions. In an attempt to facilitate some of these listening practices (particularly percussion), we introduce a hear-through system for intra-stimulus Auditory Contrast Enhancement (ACE) in real time. Plausible spectral ACE is achieved by adopting the neural mechanism of lateral inhibition. Additional decay prolongation facilitates pitch perception. Perceptual plausibility of the augmented auditory feedback from the observer-perspective is investigated in an experiment with auditory-visual stimuli. Measured plausibility forms material-specific patterns depending on spectral dynamics and decay prolongation.

CCS CONCEPTS

• **Human-centered computing** → **Auditory feedback**; *Mixed / augmented reality*; *Sound-based input / output*; Activity centered design.

KEYWORDS

augmented auditory feedback, auditory augmentation, auditory contrast enhancement, ace, auditory display, spectral contrast, percussion, auscultation, cartoonification

ACM Reference Format:
Marian Weger and Robert Höldrich. 2019. A hear-through system for plausible auditory contrast enhancement. In *Audio Mostly (AM'19), September 18–20, 2019, Nottingham, United Kingdom.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3356590.3356593

AM'19, September 18–20, 2019, Nottingham, United Kingdom

ACM ISBN 978-1-4503-7297-8/19/09…$15.00
https://doi.org/10.1145/3356590.3356593

Figure 1: Schematic drawing of the proposed ACE hear-through system. Adapted from [14, p. 6] [Public domain].

1 INTRODUCTION

Our ears are high-performance sensor devices with remarkable capabilities. We might not always be completely aware of their valuable contribution to our daily routines; however, we incessantly adapt to the auditory feedback of physical interactions with our environment. We adjust our walking pace to the sound of the floor, we shake a gift box to guess its content, we monitor the inner status of machines such as cars by their sounds. In summary, auditory feedback keeps us in the control loop of everyday and professional practices.

A very common listening practice is to knock on an object to obtain information on its inner structure. This can be seen as a rudimentary form of an impulse response measurement. Based on our experience, the amplitudes, frequencies, and decay times of resonant modes give a hint on physical properties such as geometry, material, or the size of a hidden cavity. For example, we locate cavities and studs behind a wall by thumping in order to find an appropriate spot for a drill hole. We knock on a watermelon to tell if it is ripe. Violin makers tune the top and back plate of the instrument by holding it lightly between two fingers and tapping it, which reveals the so-called *tap tones*. Similarly, the level of wine or beer in a

wooden barrel is estimated by ear via knocking. This active listening practice inspired the son of an innkeeper and later physician Leopold Auenbrugger to his "inventum novum" [3, 19]: *percussion*. As a method of clinical examination, it has become indispensable in the repertoire of modern physicians — together with its passive counterpart, *auscultation*, which is usually done by using a stethoscope. Despite the medical definitions, percussion and auscultation are here interpreted in a broader sense where the examined object can be any physical object.

For fundamental research, it is of particular importance to rethink any established routines and to investigate visions of future tools, no matter of their technical feasibility at the moment. What if we were able to artificially modify the auditory feedback of our interactions? The resulting *augmented auditory feedback* could be derived from either the original sound or from the performed action. Such augmented auditory feedback can follow three goals. (1) Add additional information to the sound; we then speak of Auditory Augmentation [5, 9, 10, 26]. (2) Modify the information that is contained in the sound, e.g., to induce a certain change in behavior; for example, by adding the sound of deep snow to the footsteps and thus subliminally manipulate walking pace [23]. (3) Enhance the information that is contained in the sound. This we call Auditory Contrast Enhancement (ACE).

We differentiate between two types of auditory contrast. *Intra-stimulus contrast* describes the prominence of those features that characterize a single sound. As an example, for some Stradivari violins, the tap tone of the top plate exhibits a prominent resonance between $C\#_4$ and D_4 [13]. For this feature, intra-stimulus contrast can be quantified by the amplitude difference between peak and average of the spectrum. By *inter-stimulus contrast*, we mean the perceptual differences between two or more sounds or groups of sounds. Sticking to the previous example, the tap tone of the bottom plate is usually one or two semitones higher [13]. Inter-stimulus contrast between the tap tones of top and bottom plate could be quantified by their spectral differences, i.e., differences in amplitude, frequency, and bandwidth. Apart from this example, both types of auditory contrast can be regarded from different perspectives, e.g., focusing on spectral, temporal, or spectrotemporal features.

We recently introduced novel methods for the enhancement of auditory contrast. *Intra-stimulus ACE* is a method to make intrinsic features of a sound more prominent, in order to facilitate their perception and thus improve the conveyance of the underlying information [27]. This is possible in real time. The underlying information may be, for example, physical properties such as material, geometry, or size of a cavity. The outcome can be interpreted as a *cartoonification* of the original sound. *Inter-stimulus ACE* is a method for the auditory display of differences between groups of sounds, based on supervised or unsupervised learning from pre-recorded samples [11]. This requires the acquisition of training data. However, a thus produced spectral ACE filter can then be applied on an unknown audio signal in real time, to facilitate a classification by ear.

In summary, intra-stimulus ACE tries to improve absolute identification of a single sound, while inter-stimulus ACE tries to facilitate discrimination between multiple sounds. Nevertheless, both are tightly connected. In case of the tap tones, the broadband resonances of top and bottom plate differ only slightly in frequency. A bandwidth reduction, as achieved by intra-stimulus ACE, transforms them into more tonal signals with higher pitch strength, and consequently higher inter-stimulus contrast. The algorithm for intra-stimulus ACE is briefly summarized in Sec. 2.

In Sec. 3, we present a practical implementation of intra-stimulus ACE based on a microphone-earphone combination, as illustrated in Fig. 1. Used as a tool, it works similar to a hearing aid, with the goal to enhance the mentioned listening practices, with a focus on percussion. In other words, the system should help to identify or categorize short impact sounds, as well as to discriminate between them, by ear. Interpretation and decision-making is left to the user.

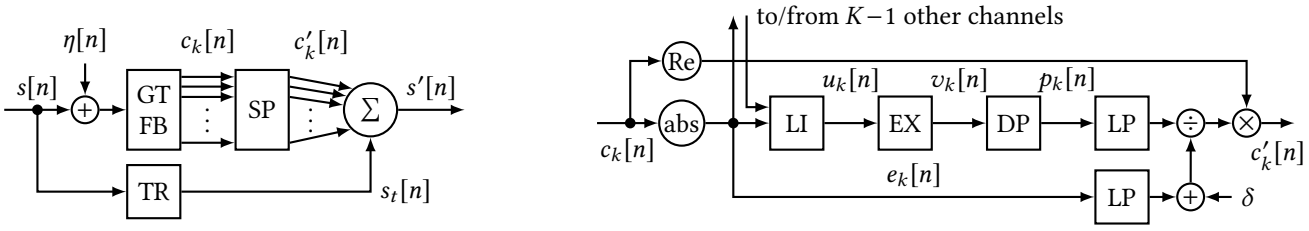
The described algorithm for intra-stimulus ACE is supposed to maintain high plausibility of the auditory feedback, as it is based on the same mechanism that is used in the neural networks of auditory nerves and auditory cortex [15, 22], namely lateral inhibition. It describes the inhibitory effect of stimulation in neighboring channels. We further use a physically meaningful decay prolongation with frequency-dependent exponential decay, in order to provide listeners with more time to perceive characteristic partials. The plausibility of the resulting auditory feedback has been evaluated in an auditory-visual experiment (Sec. 4). The main objective was to find a compromise between high auditory contrast and acceptable plausibility for observed interactions. In addition, the results should give an insight on the relation between sound cartoonification and plausibility.

2 REAL-TIME INTRA-STIMULUS ACE

Figure 2a shows the overall block diagram of the algorithm. Output $s'[n]$ is a sum of individually processed sub-band signals (spectral ACE) and a transient signal $s_t[n]$ obtained from transient restoration (TR). Spectral ACE is partly based on algorithms for speech enhancement (see [28] for an overview). Sound examples for all processing steps are available online[1].

A Gammatone Filterbank (GTFB) is used to split the input signal $s[n]$ into K complex-valued sub-bands $c_k[n]$; k is the channel index. Gammatone filters are a widely used model for the auditory filters [24]. The individual filter outputs are then

[1] ACE sound examples: https://doi.org/10.4119/unibi/2935786



(a) overall block diagram of ACE

(b) block diagram of one channel of sub-band processing (SP)

Figure 2: Block diagram for real-time Auditory Contrast Enhancement (ACE).

fed into the sub-band processing block (SP), where the actual spectral contrast enhancement takes place. The enhanced output signal is obtained by summing the processed (real-valued) sub-bands $c'_k[n]$.

We use $K = 60$ 4th-order Gammatone filters with center frequencies from 50 Hz to 20 kHz, overlapping at their -4 dB cutoff frequencies. Their bandwidths are set constant in parts of the Equivalent Rectangular Bandwidth (ERB) [18]. Center frequencies are equally spaced on the ERB-rate scale which corresponds to an equal spacing on the basilar membrane. Bandwidth in Hz decreases with decreasing center frequency, which leads to longer impulse response and group delay towards low frequencies. For resynthesis (summation of the processed sub-bands), the channels are weighted with alternating signs to achieve low ripple in the output spectrum without the need for delay compensation (see [20]).

Each sub-band channel $c_k[n]$ individually passes sub-band processing as shown in Fig. 2b. First, the sub-band envelope $e_k[n]$ is estimated by taking the absolute value of $c_k[n]$. A series of processing blocks enhances auditory contrast in different ways: spectral sharpening via lateral inhibition (LI), spectral dynamics expansion via exponentiation (EX), and decay prolongation (DP). The altered envelope $e'_k[n]$ is finally applied to the real part of the sub-band signal $c_k[n]$ by multiplication with the ratio between processed and original envelope. Both are low-pass filtered by a leaky integrator (LP) with time constant $\tau = 2$ ms to suppress disturbing artifacts which occur at high amplitude ratios, especially at low overall volume. A small value $\delta = 10^{-5}$ is added to the denominator for regularization.

Spectral sharpening via Lateral Inhibition

Spectral sharpening can be seen as the auditory counterpart to edge detection in the visual domain. It is based on lateral inhibition, i.e., convolution with a Difference-of-Gaussians (DoG) function. The inhibitory influence of neighboring bands on the observed one decreases with growing distance, following a Gaussian function on the ERB-rate scale. Bandwidth is set in terms of standard deviation σ in ERB-rate. Edge effects at the lowest and highest sub-band (cf. [15]) are suppressed by introducing two virtual sub-bands (copies of sub-bands 2 and $K-1$) as sub-bands 0 and $K+1$.

The inhibition term $T_k[n]$, as defined in Eq. 1, quantifies the overall energy in the neighboring sub-bands. Basing its calculation on the low-pass filtered sub-band envelopes $\tilde{e}_k[n]$ (leaky integrator with time constant τ_{LI}) suppresses inhibition caused by short spikes in neighboring bands, and adds an aftereffect:

$$T_k[n] = \sqrt{\frac{1}{2\gamma_k^-}\sum_{i=0}^{k-1}\gamma_{i,k}\cdot\tilde{e}_i^2[n] + \frac{1}{2\gamma_k^+}\sum_{i=k+1}^{K+1}\gamma_{i,k}\cdot\tilde{e}_i^2[n]}\,, \quad (1)$$

where $\gamma_{i,k}$ is a Gaussian function, with center frequencies f_c of the filters given in ERB-rate:

$$\gamma_{i,k} = \exp\left(-(f_{c,i} - f_{c,k})^2/(2\sigma^2)\right)\,. \quad (2)$$

The scaling factor is omitted, as the weights are anyway normalized for lower and upper neighbors individually:

$$\gamma_k^- = \sum_{i=0}^{k-1}\gamma_{i,k} \qquad \text{and} \qquad \gamma_k^+ = \sum_{i=k+1}^{K+1}\gamma_{i,k}\,. \quad (3)$$

With spectral sharpness controlled by parameter $\rho \geq 0$, the sharpened envelopes $u_k[n]$ then become:

$$u_k[n] = e_k[n]\cdot\min\left\{\left(\frac{\tilde{e}_k[n]}{T_k[n]}\right)^\rho\,,\ 1\right\}\,. \quad (4)$$

Spectral Dynamics Expansion via Exponentiation

Spectral dynamics expansion is achieved by exponentiation of the magnitude spectrum. Likewise spectral sharpening, it operates on the smoothed sub-band envelopes $\tilde{u}_k[n]$ (leaky integrator with time constant τ_{EX}). Envelopes $u_k[n]$ are scaled with respect to the global maximum $\tilde{u}_{max}[n]$, exponentiated by a parameter $\beta \geq 0$, clipped, and scaled back:

$$v_k[n] = u_k[n]\cdot\min\left\{\left(\frac{\tilde{u}_k[n]}{\mu\tilde{u}_{max}[n]}\right)^\beta\,,\ \frac{\tilde{u}_{max}[n]}{\tilde{u}_k[n]}\right\}\,, \quad (5)$$

where $\tilde{u}_{max}[n]$ is the instantaneous global maximum of $\tilde{u}_k[n]$, and μ is a threshold parameter.

Decay Prolongation

Natural interaction sounds often consist of a short transient sound with high bandwidth and a number of resonant frequencies with exponential decay. Only the decay part receives decay prolongation; both are re-combined afterwards.

We use two non-linear low-pass filters based on a leaky integrator: $\mathrm{env_a}$ with smooth attack but instant decay, and $\mathrm{env_d}$ with smooth decay but instant attack. $\mathrm{env_a}$ is given in Eq. 6 for an arbitrary input $x[n]$ and output $y[n]$. For $\mathrm{env_d}$, the same equation applies, however, with flipped direction of the inequality sign, leading to exponential decay.

$$y[n] = \begin{cases} (1-\alpha)|x[n]| + \alpha y[n-1], & |x[n]| < y[n-1] \\ |x_k[n]|, & \text{otherwise} \end{cases} \quad (6)$$

Smoothing factor α is parametrized via time constant τ or -60 dB reverberation time T_{60} ($\alpha = \exp(-T_s/\tau)$, T_s is the sampling interval, $\tau = T_{60}/\ln(1000)$).

The envelope with smoothed attack ($\mathrm{env_a}\{v_k[n]\}$) is fed to decay prolongation, while the residuum ($e_k[n]$–$\mathrm{env_a}\{v_k[n]\}$) containing only the attack part is added back to the result:

$$p_k[n] = \mathrm{env_d}\Big\{\mathrm{env_a}\big\{v_k[n]\big\}\Big\} + v_k[n] - \mathrm{env_a}\big\{v_k[n]\big\} \,. \quad (7)$$

Decay prolongation is fed by intrinsic signal components. In case of a large Signal-to-Noise Ratio (SNR) combined with long decay prolongation, a pink noise signal $\eta[n]$ (-96 dBFS is sufficient) is added to the input signal just before feeding it to the Gammatone filterbank (see block diagram in Fig. 2a). The control parameter for decay prolongation is T_{60} at 1 kHz. The decay time for each sub-band is set inversely proportional to its center frequency; it is clipped below 1 kHz.

Transient Restoration

Even if sub-band processing is bypassed, the filterbank itself induces a frequency-dependent group delay which is small at high frequencies but increases towards lower frequencies (23.5 ms for the lowest band at 50 Hz). This might be acceptable for steady sounds; however, in the extreme case, any short transient is transformed into a down-chirp. Transients are further smeared by spectral sharpening and spectral dynamics expansion. For maximum spectral contrast in addition with instantaneous and sharp transients, they must be detected as fast as possible from the original input signal, and mixed to the output of spectral ACE.

Transients are detected by the same simple transient detection algorithm that is used for decay prolongation. A 2nd-order high-pass filter with adjustable cutoff frequency shifts its focus on high-frequency content. The envelope $e_t[n]$ of the transient is estimated via the difference of a slowly decaying envelope $e_{t,d}[n]$ and a slowly rising envelope $e_{t,a}[n]$:

$$e_t[n] = \max\left\{e_{t,d}[n] - e_{t,a}[n] - \nu \,,\ \ 0\right\} \,, \quad (8)$$

with threshold ν. Envelopes are computed via the two filters $\mathrm{env_d}$ and $\mathrm{env_a}$ from Eq. 6, with time constants τ_d and τ_a, respectively:

$$e_{t,d}[n] = \mathrm{env_d}\big\{s_h[n]\big\} \quad , \quad e_{t,a}[n] = \mathrm{env_a}\big\{e_{t,d}[n]\big\} \,, \quad (9)$$

where $s_h[n]$ is the input signal $s[n]$, high-pass-filtered at f_c. The output signal $s_t[n]$ contains only the detected transients with their original amplitude:

$$s_t[n] = s[n] \cdot e_t[n] \,/\, \mathrm{env_d}\big\{e_t[n]\big\} \,. \quad (10)$$

3 THE ACE HEAR-THROUGH SYSTEM

The technical setup of the ACE hear-through system is based on earphones with integrated binaural microphones, as illustrated in Fig. 1. We use the Roland CS-10EM microphone/earphone combination. The software is implemented in SuperCollider[2] and should therefore run on most available platforms. It is published as free software via a public code repository[3]. We build upon the Gammatone filter design by Hohmann [12]; its SuperCollider implementation[4] has been adapted to output real and imaginary part of the signal. The system provides a reduced GUI which should be sufficient for most tasks. However, expert users are always able to switch to an extended GUI which allows tuning of all the described parameters of spectral and temporal ACE.

For auditory-tactile environments as well as for hearing aids, a round-trip latency of less than 10 ms is usually recommended [1, 6]. This is hard to achieve on a standard laptop; we therefore use a workstation PC with RME HDSPe MADI FX audio interface for the first prototype. Measured round-trip latency of this setup, from the loudspeaker of one earpiece to its integrated microphone, is 3.6 ms (at 48 kHz sampling rate). Note, however, that the frequency-dependent group delay of the filterbank must be added to this value: a latency below 10 ms is achieved only for the 19th band (802 Hz) and above. Below that band, the summed latency grows up to 27.1 ms (at 50 Hz).

The system does not alter inter-aural time differences, but inter-aural level differences might be altered quite unpredictable, if both channels are processed individually. As localization is anyway bad with such hear-through systems [17], we decided to sum both microphone signals to a monophonic signal before being processed through the algorithm for real-time intra-stimulus ACE. The same output signal is then played back by both earphones.

Demo videos with different materials are available online[5].

[2] SuperCollider: https://supercollider.github.io/
[3] ACE public code repository: https://git.iem.at/weger/ace
[4] *Gammatone* UGen in *AuditoryModeling* UGens in *SC3 Plugins*: https://github.com/supercollider/sc3-plugins
[5] ACE hear-through system demo videos: http://phaidra.kug.ac.at/o:91924

Table 1: Properties of the audio/video files used in the experiment: material and object category, original filename, starting time in the recording in s, and acoustical descriptors, i.e., Spectral Center of Gravity (SCG) in kHz, Spectral Bandwidth (SB) in kHz, and average −60 dB decay time RT_{60} in s.

ID	material/object	filename prefix	time	SCG	SB	RT_{60}
0	cardboard box	2015-03-30-01-38-59	1.60	2.66	3.81	0.33
1	glass table	2015-03-30-01-29-03	9.31	7.86	5.09	0.45
2	ceramic cup	2015-02-16-16-49-06	3.24	6.89	4.25	0.28
3	forest soil	2015-09-23-16-13-51-446	21.66	3.03	4.56	0.27
4	plastic box	2015-03-30-01-54-29	2.50	2.77	3.34	0.47
5	plastic bottle	2015-03-30-02-10-02	8.81	2.69	3.50	0.34
6	wooden bar	2015-10-06-18-02-12-1	1.00	4.71	4.31	0.14
7	metal box	2015-03-30-02-18-31	0.50	4.84	4.76	0.38
8	glass bowl	2015-03-25-00-09-47	18.95	5.86	4.03	0.65
9	metal table	2015-03-30-02-22-11	19.32	7.28	4.38	0.31

4 PLAUSIBILITY OF ENHANCED AUDITORY FEEDBACK

Like any other interface, the ACE hear-through system might not be accepted by the users if it lacks a plausible or natural feel [25]. We obviously need to find a compromise between auditory contrast and plausibility of the augmented auditory feedback. Therefore, we designed an experiment to evaluate the perceived plausibility of the auditory feedback of physical interactions from the observer-perspective. Participants watched short video sequences and were asked to rate the plausibility of the (augmented) audio track.

Stimuli

The total 260 stimuli for the experiment consist of 10 videos of 3 seconds duration, each with 26 alternative audio tracks. Recordings are taken from the Greatest Hits dataset[6] [21], a collection of audio/video recordings of different kinds of objects and materials being hit with a drumstick.

Selection. The selection of adequate audio/video sequences from the Greatest Hits dataset was based on several subjective and objective requirements. (1) Objects are rigid and stable (no water, leaves, gravel, etc.). (2) There is no clipping in the audio recording. (3) There should be a variability of physical properties (material, geometry, etc.). (4) There should be a variability of sound properties (pitch, timbre, decay time, etc.). (5) There is a period of 3 seconds which contains several differently sounding hits.

We initially chose 15 stimuli from which we eliminated three that produced redundant data points in the 3D parameter space of spectral centroid, spectral bandwidth, and average decay time (see Tab. 1) [2]. Additional two stimuli were excluded, as they led to very saturated results in a first pilot test (either always plausible or always implausible). Table 1 lists the remaining 10 stimuli with additional data such as material and type of the object as well as corresponding name and starting time in the dataset.

Pre-processing. The original (already synchronized) audio and video tracks were cut frame-aligned (29.97 fps) with FFmpeg[7]. Videos were re-encoded to 1080p MJPEG, in order to minimize computational effort during playback. Only the first of the two audio channels was used.

A noise sample was taken from a 'silent' region of each stimulus. It was used as a profile for the built-in noise reduction of Audacity[8] (3 bands frequency smoothing, sensitivity of 6) to reduce background noise of the stimulus by −20 dB.

Spectral ACE. Spectral dynamics are defined by ρ (sharpening) and β (expansion). We selected five pairs of values which form clearly perceivable steps from very gentle sharpening up to extreme dynamics expansion. Values are $\rho = \{2, 6, 25, 25, 25\}$, and $\beta = \{0, 0, 0, 1, 9\}$, respectively; they have been chosen in informal listening sessions. The selected decay times are equally spaced on a logarithmic scale, which is an adequate approximation of their perception [4]: $T_{60} = \{0, 0.15, 0.36, 0.84, 2\}$ (in seconds). In addition to all 25 combinations of spectral dynamics and T_{60}, there was a control condition with bypassed spectral ACE ($\rho = \beta = T_{60} = 0$, but still passing the filterbank). The original, unprocessed recording was not used, to prevent participants from identifying it as a reference. Remaining parameters were set to fixed values: $\sigma = 3$ ERB, $\tau_{LI} = \tau_{EX} = \tau_{DP} = 7$ ms.

Resynthesis and Transient Restoration. The noise sample from pre-processing was again used to re-synthesize a similar noise floor in Matlab — with identical magnitude spectrum (and thus level), but with random phase spectrum. This noise is added to the output of ACE, to provide an ecologically valid scenario, circumventing the by-product of noise reduction that comes through spectral ACE. Parameters for transient restoration are set to $f_c = 4$ kHz, $\nu = -42$ dB, $\tau_d = 60$ ms, and $\tau_a = 3$ ms; the transient is added with −3 dB gain.

Participants, Apparatus, and Procedure

Eleven participants (4 female, 7 male) were recruited from university staff. Those were happy to take part without payment. They all have a professional background in sound and music computing and are thus regarded as expert listeners.

The experiment software was implemented in Pure Data[9], with GUI and video playback realized in Gem[10]. Auditory stimuli were pre-rendered in Matlab. We used a standard laptop computer with 1920×1080 pixels screen, running Debian Linux. Sound was presented via AKG K 272 HD closed-back headphones which were directly connected to the integrated Conexant CX8200 audio codec.

[6]The Greatest Hits dataset: http://camouflage.csail.mit.edu/vis/

[7]FFmpeg: http://ffmpeg.org/

[8]Audacity: https://www.audacityteam.org/

[9]Pure Data: https://puredata.info/

[10]Gem: http://gem.iem.at/

The 260 stimuli were presented without repetitions; the order was randomized between participants. Randomization was constructed of 26 consecutive rows, each including all 10 videos in random order; with the restriction that the first video of a row differs from the last video of the previous row. For each video individually, the 26 audio tracks were randomly distributed to the 26 appearances of the video.

In each trial, the auditory-visual stimulus was automatically played back once. Participants were able to replay the stimulus in case of distraction. They were asked to rate the audio track with respect to its plausibility for the shown video. The answer had to be given on a 5-point scale, ranging from 'very implausible' (scale value 0) over 'borderline' (0.5) to 'very plausible' (1). The in-between values (0.25 and 0.75) had no label. To speed up the experiment, the answer could already be given before the end of the video. The next trial started automatically, 600 ms after giving the answer.

Prior to the experiment, participants had to read a short summary of this procedure, including some clarifications: (1) Sound comes from an unprofessional microphone recording. (2) The task is *not* to rate authenticity or to identify the original recording. (3) The task is to quantify the plausibility of different more or less plausible alternatives. (4) Synonyms for the admittedly vague term 'plausible' are 'natural', 'convincing', and 'realistic'. One trial was randomly drawn from the whole population of trials, to serve as demonstration of the experiment interface before start. The whole experiment took about 25 minutes.

Results

From the plausibility ratings, we obtained a frequency (number of answers) for each of the five scale values and for every stimulus. We are not able to assign numbers to the plausibility ratings, as we do not know the scaling of this abstract value — we must even assume that the mental model of plausibility is different for each participant. The obtained values are therefore interpreted as purely ordinal data.

A visual inspection of the frequencies for the scale values of all 260 stimuli did not point out any multimodal distributions; we therefore assume a unimodal distribution which validates the computation of medians. These medians (between 0 and 1) are shown in Fig. 3 for all combinations of spectral dynamics and decay time, for the individual videos.

We consider stimuli as 'acceptable' if they reach at least 'borderline' plausibility. This is the case, when the median is at least 0.5. We tested each median with a one-tailed sign test. For medians below 0.5, we tested against the null hypothesis that it is at least 0.5. For medians above or equal to 0.5, we tested against the null hypothesis that it is below or equal to 0.375. Those values where the null hypothesis could be rejected at a significance level of 0.05 are highlighted in Fig. 3.

It is clear from this data that the independent factors spectral dynamics, decay time, and video (i.e., object/material) affect the perceived plausibility. As expected, higher values of ρ/β and T_{60} generally led to lower plausibility. The control condition was always significantly above borderline plausibility. The extreme values (top row and rightmost column) are mostly below borderline plausibility. Some videos formed very specific patterns in the 2D parameter space of ρ/β and T_{60}. The cardboard and plastic objects achieved the lowest plausibility ratings. The cardboard box was never significantly above borderline plausibility, except in the control condition. For the plastic objects, only small values of ρ/β and T_{60} were accepted. The metal box achieved highest plausibility ratings within the main diagonal, when both parameters are increased together. For ceramic, soil, and wood, only the decay prolongation seems to have a significant effect on perceived plausibility. The glass and metal objects were almost always plausible, except for extreme values.

These observations led us to do a cluster analysis, in order to validate how the individual results of the videos group together. We performed a hierarchical cluster analysis based on the pooled median values for each video. Independent of distance metrics (inner squared distance, farthest distance, shortest distance) or data (medians, means, % answers ≥ 0.5), two main clusters are emerging very clear. Cluster A contains the four metal and glass objects, while Cluster B includes wood, soil, cardboard, plastic, and ceramics. The resulting median values of these clusters are shown in the bottom right of Fig. 3. For metal and glass, only the highest value of spectral dynamics and decay, respectively, is rated as implausible. The highest combination with acceptable plausibility is $\rho = 25$ / $\beta = 1$ / $T_{60} = 0.84$ s (the extreme value of $T_{60} = 2$ s is considered to be too large, even if it is still rated with borderline plausibility). Cluster B obviously incorporates the same tendencies as the individual videos for wood, ceramics, plastic, cardboard, and soil: low overall plausibility, weak impact of spectral contrast, and a preference for the diagonal. Highest acceptable values are $\rho = 25$ / $\beta = 0$, $T_{60} = 0.15$ s.

Discussion

Some of the above results might allow us to draw conclusions on the participant's mental model of plausibility in the context of the experiment. In general, we consider sensory feedback as plausible if it is "conceptually consistent with what is known to have occurred in the past" [7]. "A highly plausible scenario is one that fits prior knowledge well: with many different sources of corroboration, without complexity of explanation, and with minimal conjecture" [7].

The cluster analysis clearly divides the stimuli in high-density materials (glass and metal) and low-density materials (plastic, wood, ceramics, soil, and cardboard). Such grouping into gross density categories has already been observed in

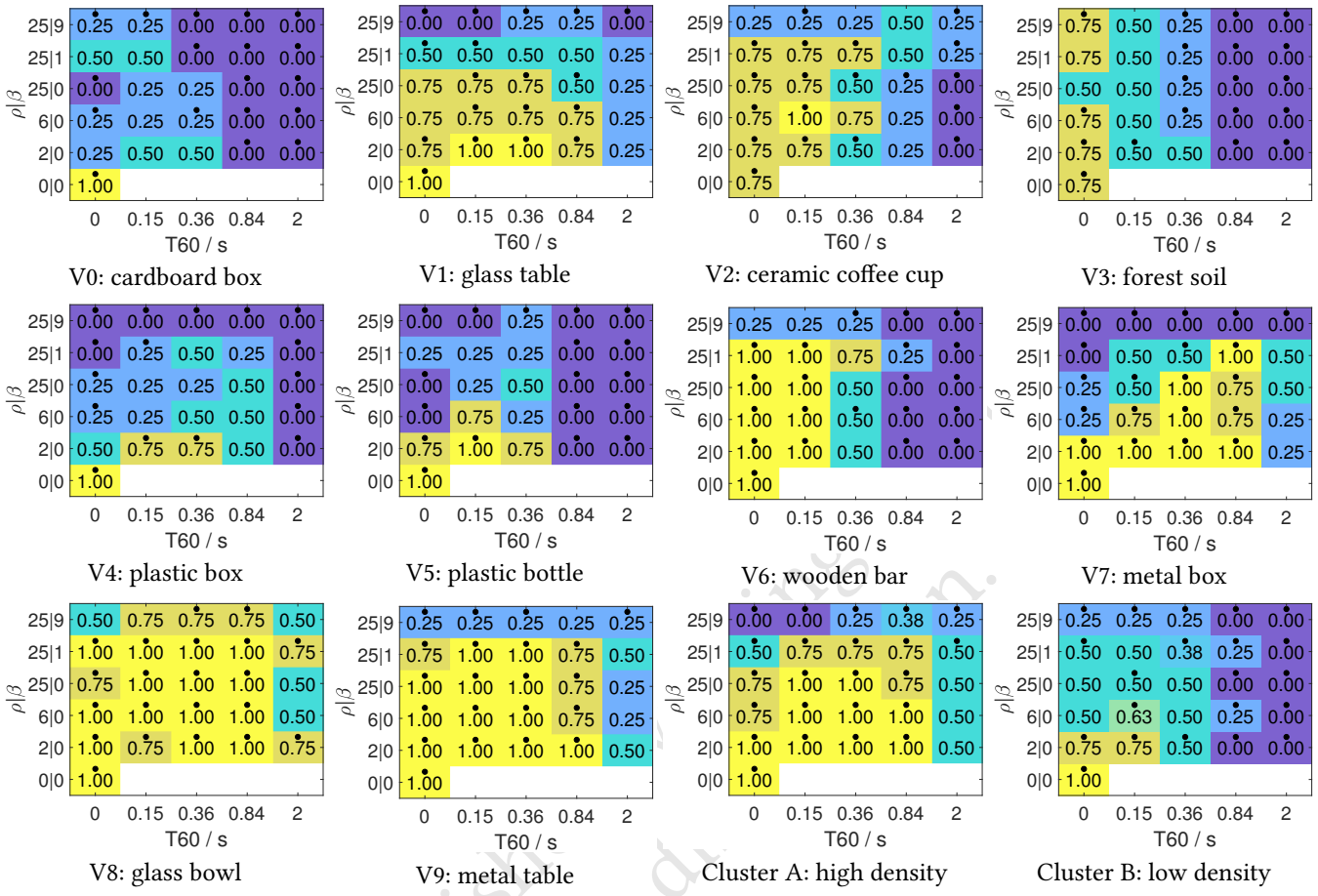


Figure 3: Experiment results. Median plausibility of the 10 videos and their two main clusters, for all 26 combinations of spectral dynamics (ρ/β) and decay time T_{60} at 1 kHz, respectively. Values reach from 0 (very implausible) to 1 (very plausible). Values that are either significantly below 0.5 or significantly above or equal to 0.5 (significance level 0.05) are marked by a dot.

other listening experiments from the literature; people can hardly discriminate materials within those groups [8].

Within the high-density materials, the plausibility of the metal table and glass bowl was almost unaffected by ACE, as these exhibit anyway strong spectral sharpness and long decay. Plausibility suffered only if disturbing artifacts occurred with extreme values of spectral dynamics (jumps between partials) and long decay time (sharp partials transform into bandpass noise). Interestingly, the glass bowl was partly rated above borderline with the 2 s decay prolongation, although it is clearly visible that it is placed on its highly damping plastic cover. This fact was easily ignored by the participants. The metal box is actually a paper dispenser which is included in two versions (filled and empty) in the Greatest Hits dataset. We selected the empty one; however, its inner structure and content is not completely visible, which might be an explanation for its insensitivity to decay prolongation. In a physically plausible scenario, this would also come together with increased spectral dynamics.

Within the low-density materials, the plastic bottle and cardboard box achieved exceptionally low plausibility ratings. An explanation might be the low frequency of the strongest resonance. For the cardboard box, it is at 72 Hz; the corresponding filter (f_c = 71 Hz) has already a group delay of 21.9 ms — unacceptable for plausible auditory feedback. With increasing spectral contrast, this delay gets more salient. For the forest soil, it is obvious that the participants could not find a physical explanation for a long decay time above 0.1 s. However, the soil was insensitive to spectral ACE, as even the original recording exhibits distinct pitch at different positions.

An objective measure of spectral contrast can be obtained via the entropy-based measure of spectral flatness *SF* [16]. Spectral contrast is then $SC = 1 - SF$. Overall spectral contrast $\overline{SC}$ is derived from N 50 % overlapping hann-windowed

signal blocks of 1024 samples. Spectral contrast SC_n of each block n is weighted with its energy E_n, respectively:

$$\overline{SC} = \frac{1}{\sqrt{N}} \frac{\sum_n \sqrt{E_n} SC_n}{\sqrt{\sum_n E_n}} \,. \tag{11}$$

Compared to the control condition with bypassed ACE, measured spectral contrast increases monotonically for increasing ρ/β (averaged over video and decay; from 51 % to 87 %) as well as for increasing T_{60} (averaged over video and spectral dynamics; from 38 % to 107 %). While maintaining at least borderline median plausibility (significantly ≥ 0.5), the average spectral contrast is enhanced by 49 % for cluster A and 26 % for cluster B. This confirms that auditory contrast can be plausibly enhanced by the proposed ACE method.

5 CONCLUSIONS AND OUTLOOK

We presented a hear-through system for intra-stimulus auditory contrast enhancement. It is intended to be used wherever auditory feedback is part of a knowledge-making process. The signal is enhanced in terms of spectral dynamics and decay time. Both enhancements have a positive effect on measured spectral contrast. Perceptual plausibility of the resulting sounds was evaluated in an experiment with auditory-visual stimuli. For 9 the 10 tested sounding objects, measured auditory contrast could be increased while maintaining the amount of plausibility that is necessary for natural interaction. Besides technical applications as mentioned in the introduction, we see great potential also in sound cartoonification and sound transformations for artistic purpose.

Future work will include a user study to evaluate the system in a full-modal real-time scenario that simulates its primary target application: percussion. It is further planned to optimize the implementation so that it runs on mobile devices, e.g., as a smartphone app. A fusion with the powerful non-real-time methods for inter-stimulus ACE [11] into a single mobile artifact would create a versatile tool for spontaneous journeys in sound.

REFERENCES

[1] M Ercan Altinsoy. 2012. The quality of auditory-tactile virtual environments. *JAES* 60, 1/2 (2012), 38–46.

[2] Mitsuko Aramaki, Mireille Besson, Richard Kronland-Martinet, and Sølvi Ystad. 2008. Timbre perception of sounds from impacted materials: behavioral, electrophysiological and acoustic approaches. In *Int. Symp. on Computer Music Modeling and Retrieval*. Springer, 1–17.

[3] Johann Leopold Auenbrugger. 1761. *Inventum Novum ex Percussione Thoracis Humani Ut Signo Abstrusos Interni Pectoris Morbos Detegendi.*

[4] Matthew G Blevins, Adam T Buck, Zhao Peng, and Lily M Wang. 2013. Quantifying the just noticeable difference of reverberation time with band-limited noise centered around 1000 Hz using a transformed up-down adaptive method. (2013).

[5] Till Bovermann, René Tünnermann, and Thomas Hermann. 2010. Auditory augmentation. In *Innovative Applications of Ambient Intelligence: Advances in Smart Systems*. IGI, 98–112.

[6] Lars Bramsløw. 2010. Preferred signal path delay and high-pass cut-off in open fittings. *Int. Journal of Audiology* 49, 9 (2010), 634–644.

[7] Louise Connell and Mark T Keane. 2006. A model of plausibility. *Cognitive Science* 30, 1 (2006), 95–120.

[8] Bruno L Giordano and Stephen McAdams. 2006. Material identification of real impact sounds: Effects of size variation in steel, glass, wood, and plexiglass plates. *JASA* 119, 2 (2006), 1171–1181.

[9] Katharina Groß-Vogt, Marian Weger, and Robert Höldrich. 2018. Exploration of Auditory Augmentation in an Interdisciplinary Prototyping Workshop. In *Forum Media Technology*.

[10] Katharina Groß-Vogt, Marian Weger, Robert Höldrich, Thomas Hermann, Till Bovermann, and Stefan Reichmann. 2018. Augmentation of an institute's kitchen: An ambient auditory display of electric power consumption. In *ICAD*.

[11] Thomas Hermann and Marian Weger. 2019. Data-driven Auditory Contrast Enhancement for everyday sounds and sonifications. In *ICAD*.

[12] Volker Hohmann. 2002. Frequency analysis and synthesis using a Gammatone filterbank. *Acta Acustica united with Acustica* 88, 3 (2002), 433–442.

[13] Carleen Maley Hutchins. 1962. The physics of violins. *Scientific American* 207, 5 (1962), 78–93.

[14] Boston Industrial School Association. 1881. *Wood-working Tools: How to Use Them. A Manual.* Ginn & Heath, for the Industrial School Association. https://archive.org/details/woodworkingtool00bostgoog

[15] A Kral and V Majernik. 1996. On lateral inhibition in the auditory system. *General physiology and biophysics* 15 (1996), 109–128.

[16] Nilesh Madhu. 2009. Note on measures for spectral flatness. *Electronics letters* 45, 23 (2009), 1195–1196.

[17] Georgios Marentakis and Rudolfs Liepins. 2014. Evaluation of hear-through sound localization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 267–270.

[18] Brian CJ Moore and Brian R Glasberg. 1996. A revision of Zwicker's loudness model. *Acta Acustica utd. with Acustica* 82, 2 (1996), 335–345.

[19] Sherwin B. Nuland. 2005. *Doctors: The History of Scientific Medicine Revealed Through Biography.* The Teaching Company.

[20] Martin Opitz, Robert Höldrich, Franz Zotter, and Markus Noisternig. 2009. Method and device for low-latency auditory model-based single-channel speech enhancement.

[21] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. 2016. Visually indicated sounds. In *IEEE Conference on CV and Pattern Recognition*. 2405–2413.

[22] C Pantev, H Okamoto, B Ross, W Stoll, E Ciurlia-Guy, R Kakigi, and T Kubo. 2004. Lateral inhibition and habituation of the human auditory cortex. *European Journal of Neuroscience* 19, 8 (2004), 2337–2344.

[23] Stefano Papetti and Federico Fontana. 2012. Effects of audio-tactile floor augmentation on perception and action during walking: Preliminary results. In *Sound and Music Comp. Conference (SMC)*. 17–22.

[24] RD Patterson, Ian Nimmo-Smith, John Holdsworth, and Peter Rice. 1987. An efficient auditory filterbank based on the gammatone function. In *Meeting of the IOC Speech Grp. on Auditory Modelling at RSRE*, Vol. 2.

[25] Patrick Susini, Nicolas Misdariis, Guillaume Lemaitre, and Olivier Houix. 2012. Naturalness influences the perceived usability and pleasantness of an interface's sonic feedback. *Journal on Multimodal User Interfaces* 5, 3-4 (2012), 175–186.

[26] Marian Weger, Thomas Hermann, and Robert Höldrich. 2018. Plausible auditory augmentation of physical interaction. In *ICAD*.

[27] Marian Weger, Thomas Hermann, and Robert Höldrich. 2019. Real-time Auditory Contrast Enhancement. In *ICAD*.

[28] Jun Yang, Fa-Long Luo, and Arye Nehorai. 2003. Spectral contrast enhancement: Algorithms and comparisons. *Speech Communication* 39, 1-2 (2003), 33–46.