REAL-TIME AUDITORY CONTRAST ENHANCEMENT

Marian Weger[1], *Thomas Hermann*[2], *Robert Höldrich*[1]

[1] IEM, University of Music and Performing Arts, Graz, Austria
[2] Ambient Intelligence Group, CITEC, Bielefeld University, Bielefeld, Germany
weger@iem.at

ABSTRACT

Every day, we rely on the information that is encoded in the auditory feedback of our physical interactions. With the goal to perceptually enhance those sound characteristics that are relevant to us — especially within professional practices such as percussion and auscultation — we introduce the method of real-time Auditory Contrast Enhancement (ACE). It is derived from algorithms for speech enhancement as well as from the remarkable sound processing mechanisms of our ears. ACE is achieved by individual sharpening of spectral and temporal structures contained in a sound while maintaining its natural gestalt. With regard to the targeted real-time applications, the proposed method is designed for low latency. As the discussed examples illustrate, it is able to significantly enhance spectral and temporal contrast.

1. INTRODUCTION

Every sound that we encounter in our daily lives contains information. If the sound is the result of a physical process such as an interaction with our environment, then it contains information on the involved physical objects (e.g., material or geometry), their environment (e.g., room acoustics), and the type of interaction (e.g., hitting or scratching). Pieces of information that are not only restricted to natural sounds but also apply for synthesized sounds are, for example, sound parameters such as frequency or amplitude, as well as their perceptual pendants — here pitch and loudness. If such sound parameters are deliberately modified with respect to some underlying data, as being the case in auditory display and also in music, then even this data is encoded in the sound. Unfortunately, we are not able to perceive the entire information, but only a small fraction of it.

Nevertheless, as an everyday experience, we rely on the auditory feedback of our physical interactions, either consciously, e.g., when shaking a box to guess its contents, or unconsciously, when automatically adapting to the physical structure of the ground while walking. If the auditory feedback (the sonic reaction to physical interaction) is artificially modified, then we speak of *augmented auditory feedback* [1]. It seeks to attain three goals. (1) Add additional information to the sound. This is usually referred to as *Auditory Augmentation* [1, 2, 3, 4]. (2) Modify the information that is already contained in the sound, in order to achieve a change in behavior, e.g., [5, 6, 7]. (3) Enhance the information that is already contained in the sound, e.g., improvement of the Signal-to-Noise-Ratio (SNR).

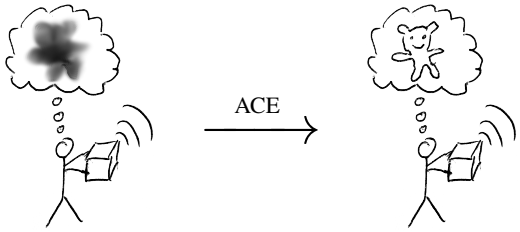


Figure 1: Someone shaking a box to guess its contents from the resulting sound. A task we want to facilitate.

In this sense, we introduce Auditory Contrast Enhancement (ACE) with the objective to enhance relevant sound characteristics in order to facilitate their perception and hence improve the conveyance of the underlying information. This concept is illustrated in Fig. 1. What might be relevant to users, however, depends on their individual activities, as well as on the type and origin of the observed sound. We expect high potential for auditory contrast enhancement where listening is part of a knowledge-making process. Especially when, for example, scientists, engineers, or physicians rely on their ears during professional routines. Even for this limited group of people and their audition-based practices, Supper and Bijsterveld discriminate between at least six different listening modes, depending on the purpose and on the way of listening [8].

One of these practices is *percussion*, a technique where a physical object or body part is actively hit in order to reveal information on its inner structure through the induced auditory feedback. This technique has established in everyday life to locate a good spot for a drill hole in a wall. The passive complement is *auscultation* where a physical object such as a machine or a human body is inspected by passively listening to its sound — usually by using a stethoscope. This tool enhances auditory contrast not only by efficient guidance of the structure-borne sound to the user's ears, but also by amplification of frequency-ranges which are of special interest to the user [9].

We distinguish between two types of auditory contrast. By *inter-stimulus contrast*, we mean the perceived differences between stimuli, which results from juxtaposing them. Inter-stimulus ACE tries to display all aspects in which two or more stimuli differ auditorily. This topic is extensively investigated in our companion paper [10] and will not be covered further here. By *intra-stimulus contrast*, we mean the strengths of peculiarity of a single stimulus. These may be the spectro-temporal dynamics of a sound. By intra-stimulus ACE, we seek to intensify those peculiarities.

https://doi.org/10.21785/icad2019.026

Our goal is to enhance the perception of those sound properties that characterize a sound, while maintaining its original gestalt as good as possible. We assume that this compromise can be achieved by attenuating non-characteristic aspects of the signal, thus leading to reduced spectral, temporal, and informational masking. In the extreme case, a very strong contrast enhancement leads to a cartoonification of the sound, reducing it to only a few very prominent sound attributes. This is conceptually similar to the visual domain where contrast is usually understood as the degree to which areas of an image differ in appearance.

Assuming that a sound is characterized by its unique spectral and temporal structure, an enhancement of this structure may automatically enhance the contrast to other sounds which exhibit a different structure. If, however, two sounds share the same strong characteristics with only minor differences, intra-stimulus contrast enhancement could even suppress those differences, leading to reduced inter-stimulus contrast between both. Such "similarity enhancement" might be useful when searching for similarities between stimuli. Otherwise, inter-stimulus contrast enhancement would be the recommended choice (see companion paper [10]).

In summary, we identify two activities which intra-stimulus ACE should improve: (1) identify the physical sound source, as visualized in Fig. 1. and (2) discriminate between sounds that are different to each other.

The rest of this article is structured as follows. In Sec. 2 we derive an algorithm for real-time intra-stimulus ACE. Spectral and temporal contrast enhancement are individually addressed in Sec. 2.1 and 2.2, respectively. Finally, a general discussion (Sec. 3) as well as conclusions and an outlook on future investigations (Sec. 4) are given. Supplementary material such as the sound examples (Snd.) referenced in the text can be found under the following link: `https://doi.org/10.4119/unibi/2935786`

2. AUDITORY CONTRAST ENHANCEMENT

The main applications that are envisaged for real-time ACE are percussion and auscultation — not so much for medical purposes but more for material testing by ear and auditory observation of mechanical processes such as machines. The targeted sounds therefore include transient interaction sounds and environmental sounds, but not speech or music. The focus on real-time application on auditory feedback makes a low-latency implementation necessary. Furthermore, the sounds resulting from ACE should maintain some degree of naturalness — they should stay within the limits of plausibility with reference to their individual context and the performed action. Even if ACE is only used as a technical tool, we know that "naturalness influences the perceived usability and pleasantness of an interface's sonic feedback" [11]. While development is performed in Matlab, the real-time algorithm will be implemented in SuperCollider and Pure Data to finally be able to run on smartphones or low-latency platforms such as the Bela [12]. Sound recording and playback can be done either with a contact microphone and loudspeaker, or by using a mic-through system (headphones with built-in microphones).

Figure 2 shows the overall block diagram. Output $s'[n]$ is a mix of three signals: (1) the dry input signal $s[n]$ (e.g., coming from a microphone), (2) the output $s_f[n]$ of Spectral Contrast Enhancement (SCE, see Sec. 2.1), and (3) the output $s_t[n]$ of Temporal Contrast Enhancement (TCE, see Sec. 2.2). Their individual gains are parametrized by two linear cross-fades: (1) between $s_f[n]$ and $s_t[n]$ to intuitively tune to the signal dimension of interest, and (2) between this weighted sum and the original signal (wet and dry) for overall strength of the effect.

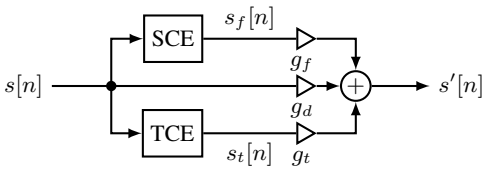


Figure 2: Overall block diagram of real-time ACE.

2.1. Spectral Contrast Enhancement

Yang et al. define spectral contrast as "the decibel difference between peaks and valleys in the [magnitude] spectrum" [13]. They describe several algorithms for spectral contrast enhancement, aiming at two applications: (1) compensation of reduced frequency selectivity in hearing-impaired people, and (2) speech enhancement in noise. One of the easiest methods is to exponentiate the magnitude spectrum by a variable exponent, followed by normalization [14]. This results in a spectral dynamics expansion with respect to the global maximum. Other approaches use linear prediction which works well for speech enhancement where detailed information on the sound source is available [13].

A large group of algorithms is based on an analog circuit proposed by Stone and Moore [15]. In principle, the signal is split into a number of frequency bands which are separately processed by a variable gain amplifier and then summed. The gain of each channel is a weighted sum of its own envelope and the envelopes of four neighboring channels; the latter with negative weights. This weighting is similar to a transversal FIR filter. As result, spectral peaks are amplified while troughs are attenuated. The digital implementation of this algorithm — Yang et al. refer to it as "Cambridge's method" — works as follows [13, 16]:

1. Computation of the spectrum X_k of a (windowed) signal block via Fast Fourier Transform (FFT), with frequency index k.
2. Calculation of excitation pattern P_k — "the representation of a spectral shape in the auditory system" [15]. It resembles a smoothed version of the magnitude spectrum $|X_k|$.
3. The enhancement function E_k is the convolution of P_k with a Difference-of-Gaussians (DoG) function. This is similar to a smoothed 2nd derivative. The DoG function is the sum of a positive Gaussian and a negative Gaussian with larger (here: $2\times$) bandwidth. Convolution runs on a scale which quantifies the number of Equivalent Rectangular Bandwidths (ERB) that fit below a certain frequency — the ERB-rate scale [17].
4. The enhanced magnitude spectrum $|Y_k|$ is then

$$|Y_k| = P_k \cdot \big(|E_k| + 1\big)^{\mathrm{sgn}(E_k)\cdot\rho}\,, \qquad (1)$$

   where $\rho \geq 0$ controls the strength of the effect.
5. Inverse FFT of $|Y_k|$ combined with the original phase values.

While Cambridge's method did not improve speech intelligibility — neither analog nor digital — its high potential in "technical" enhancement of spectral contrast, i.e., increasing differences between peaks and valleys, is evident.

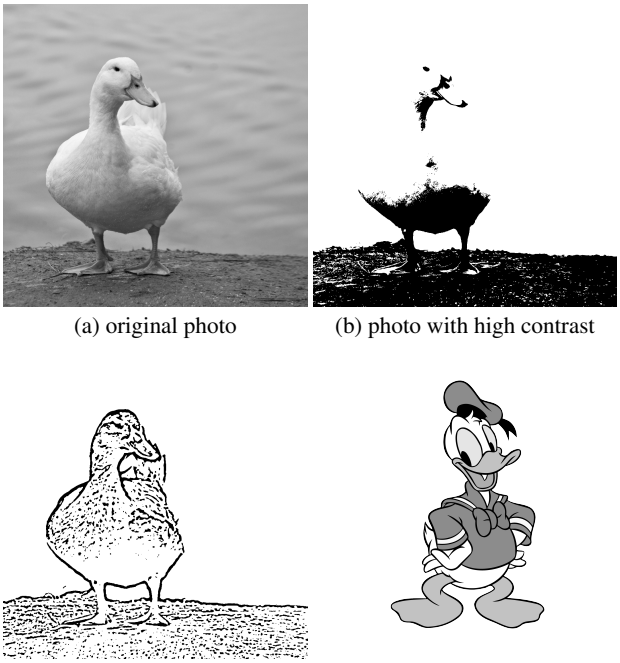
Our auditory system achieves spectral contrast enhancement similar to Cambridge's method. The underlying mechanism is

based on Lateral Inhibition (LI) in the neural networks of the auditory nerves and the auditory cortex [18, 19]. In general, this process can be described as "the suppression of nervous activity at one place in a receptor field as a consequence of the stimulation of adjacent places in this field" [20]. Besides, for instance, the retina and the skin, such receptor fields are also found along the basilar membrane [21, 22]. Kral and Majernik used an artificial neural network to model the effect of spectral contrast enhancement in the auditory system via lateral inhibition [18]. Among their simulated scenarios, three extreme cases are of particular interest. (1) Partly overlapping band-limited noise signals are narrowed in bandwidth and thus separated. (2) Uniform white noise is effectively suppressed. (3) Uniform white noise where a specific frequency-range has been suppressed leads to spikes at the edges of the stopband — the so-called edge effect.

It seems that in general there are two types of spectral contrast: (1) exponentiation relative to the global maximum (we refer to it as spectral dynamics expansion), and (2) lateral inhibition (we refer to it as spectral sharpening). It might be interesting to compare these to the visual domain. Spectral dynamics expansion compares to visual contrast control as shown in Fig. 3b, while spectral sharpening is actually edge detection (see Fig. 3c; the image shows the inverted result) — remember the edge effect demonstrated by Kral and Majernik [18]. In order to achieve something close to cartoonification, as exaggeratedly illustrated in Fig. 3d, we would need a combination of both types of contrast. In vision, this would be an overlay of Fig. 3b and c, e.g., by multiplying or taking the minimum of both images). In the auditory domain, we would take the maximum of both output spectra. The above considerations suggest that both types of spectral contrast enhancement are necessary, depending on the sound characteristics of interest, and therefore need to be implemented for parallel or serial use.

As we target low latency and real-time operation, the use of FFT — the basis for the majority of speech enhancement algorithms — is not possible. For that reason, frequency separation must be achieved by a filterbank, similar to the analog circuit by Stone and Moore [15]. We are therefore restricted to operate on a very limited number of frequency bands. Note, however, that Cambridge's method returns an altered version of the excitation patterns — a signal with significantly reduced frequency resolution. An adequate approximation of the excitation pattern can be obtained by a Gammatone filterbank (GTFB) — a widely used model for the auditory filters [23]. If the filters' center frequencies are equally spaced on the ERB-rate scale (and set to constant bandwidth in parts of the ERB), they simulate an equal spacing on the basilar membrane. The lower bands exhibit a smaller bandwidth in Hz, leading to longer impulse response and group delay. This implies a trade-off between frequency resolution and group delay towards low frequencies, which needs to be taken care of.

The excitation pattern is expressed by the energy distribution across sub-bands, calculated via their channel envelopes. Depending on the implementation of the Gammatone filter, it can also output the imaginary part of the resulting signal, in addition to the real output. An accurate estimation of the signal envelope is then given by the magnitude of the complex filter output. A suitable implementation is the one by Hohmann [24], which is available for



(a) original photo (b) photo with high contrast

(c) photo with edge detection (d) a famous cartoon duck

Figure 3: The photo of a white duck in three versions, and the drawing of a famous cartoon duck.[1]

Matlab[2], Pure Data[3] and SuperCollider[4]; in the latter case, a small modification of the source code is needed in order to return the imaginary part. We use 60 4th-order filters with center frequencies from 50 Hz to 20 kHz, overlapping at their -4 dB cutoff frequency (as in [25]). During resynthesis, i.e., summation of the processed sub-bands, their different group delays are usually compensated by individual time-delays, in order to reduce ripple in the output spectrum. We circumvent such additional latency by weighting the sub-bands with alternating signs, as proposed by Noisternig [25].

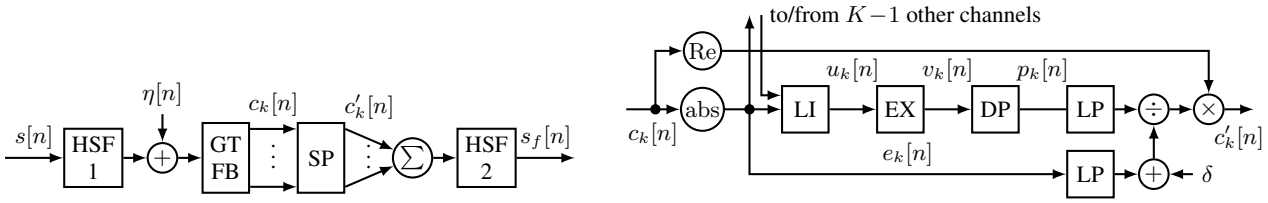
A block diagram of the proposed algorithm for spectral contrast enhancement is shown in Fig. 4. The overall block diagram (Fig. 4a) illustrates the general idea described above. In summary, the input signal $s[n]$ is split into K sub-bands $c_k[n]$ by a Gammatone filterbank with K channels; k is the channel index. The actual spectral contrast enhancement is done within the sub-band processing block (SP). The sum of the processed (real-valued) sub-bands $c'_k[n]$ then forms the enhanced output signal. Within SP, all channels are treated equally. While the Gammatone filterbank accounts for the $1/f$ proportionality of signal energy, this might not be enough for many natural signals which may exhibit even stronger high-frequency loss. This could lead to overly damped high-frequency content in the output. This effect is reduced by a pair of shelving filters (HSF) — one boosting high frequencies of

[1] Fig. 3a-c: Anne Davis, `http://flickr.com/anned/`, Creative Commons Attribution NonCommercial (CC BY-NC) 2.0 Generic License. Fig. 3d: `http://pngimg.com`, CC BY-NC 4.0 International License.

[2] Matlab implementation of the used Gammatone filterbank [24]: `http://medi.uni-oldenburg.de/download/demo/gammatone-filterbank/gammatone_filterbank-1.1.zip`

[3] *Audition library* for Pure Data: `http://lumiere.ens.fr/Audition/tools/realtime/`

[4] *AuditoryModeling* UGens from *SC3 Plugins*: `https://github.com/supercollider/sc3-plugins`



(a) overall block diagram of spectral ACE

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

Figure 4: Block diagram of spectral ACE.

$s[n]$ before feeding it to the Gammatone filterbank (HSF 1), and another one inverting the effect of the first one by attenuation after resynthesis/summation (HSF 2).

Each channel $c_k[n]$ individually passes sub-band processing as shown in Fig. 4b. First, the sub-band envelope $e_k[n]$ is extracted by taking the absolute value of the complex signal $c_k[n]$. This envelope then successively passes three stages: lateral inhibition (LI, see Sec. 2.1.1), exponentiation (EX, Sec. 2.1.2), and decay prolongation (DP, Sec. 2.1.3). The processed envelope $p_k[n]$ 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 (see Eq. 2). Both envelopes are low-pass filtered by a leaky integrator with time-constant $\tau = 2$ ms to suppress disturbing artifacts which occur at high amplitude ratios, especially at low overall volume. For regularization, a small value $\delta = 10^{-5}$ is added to the denominator (assuming audio signals in the range between -1 and 1).

$$c'_k[n] = \mathrm{Re}\{c_k[n]\} \cdot \frac{e'_k[n]}{e_k[n] + \delta} \,. \tag{2}$$

2.1.1. Spectral Sharpening

One problem we see in Cambridge's method (Eq. 1) is that it not only dampens spectral valleys but also amplifies spectral peaks. This uncontrolled amplification of the signal can be avoided by restricting the enhancement function E_k to negative values.

We first define an inhibition term $T_k[n]$ which quantifies the overall energy in the neighboring sub-bands. If it is larger than the energy in the observed band, then this band is attenuated. Calculation of the inhibition term is based on the sub-band envelopes $e_k[n]$ which are low-pass-filtered by a leaky integrator, which leads to $\tilde{e}_k[n]$. The resulting slow attack time suppresses inhibition caused by short spikes in neighboring bands, while the decay adds an aftereffect to the lateral inhibition.

We base the calculation of the neighboring bands' weights on the DoG function as in Cambridge's method. The ratio between the bandwidths of the two Gaussians controls the sharpness of the resulting spikes in the spectrum. As our approach anyway restricts sharpening to the bandwidths of the used filters (which is quite "unsharp"), we reduce the positive Gaussian to a minimum, being a Dirac delta impulse. This way, extreme enhancement (large ρ) would inhibit all frequency bands except those which describe local maxima. The bandwidth of the negative Gaussian is set via its standard deviation σ in ERB-rate.

For the lowest and highest sub-band, neighbors of significant weight are outside the scope of the filterbank. A zero-padding (insertion of zero-valued virtual bands on both sides) would introduce an unwanted edge-effect at the lowest and highest sub-band ($k = 1$ and $k = K$, respectively), similar to the simulation by Kral and Majernik [18]. Therefore, two virtual sub-bands (copies of sub-bands 2 and $K-1$) are introduced as sub-bands 0 and $K+1$, respectively (copying the edge bands themselves would half a potential contrast in those bands). The inhibition term $T_k[n]$ then becomes

$$T_k[n] = \sqrt{\frac{1}{\gamma_k^-} \sum_{i=0}^{k-1} \gamma_{i,k} \cdot \tilde{e}_i^2[n] + \frac{1}{\gamma_k^+} \sum_{i=k+1}^{K+1} \gamma_{i,k} \cdot \tilde{e}_i^2[n]} \,, \tag{3}$$

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

$$\gamma_{i,k} = \exp\left(-\frac{(f_{c,i} - f_{c,k})^2}{2\sigma^2}\right) \,. \tag{4}$$

The scaling factor can be omitted, as the weights are anyway normalized for the lower and upper neighbors individually, altogether summing up to 1:

$$\gamma_k^- = 2\sum_{i=0}^{k-1} \gamma_{i,k} \qquad \text{and} \qquad \gamma_k^+ = 2\sum_{i=k+1}^{K+1} \gamma_{i,k} \,. \tag{5}$$

This scaling ensures that a signal with equal envelopes, i.e., in which $e_k[n]$ is the same for all k, implies $T_k[n] = \tilde{e}_k[n]$, and therefore leads to unchanged envelopes. Due to the ERB-scaled Gammatone filterbank, this is the case for a pink noise signal which exhibits a magnitude spectrum that is proportional to $1/f$. This relation approximates the decrease in energy towards high frequencies, that is common to many natural sounds. In analogy to Eq. 1, the sharpened envelopes $u_k[n]$ then become

$$u_k[n] = e_k[n] \cdot \min\left\{\left(\frac{\tilde{e}_k[n]}{T_k[n]}\right)^{\rho} \,,\; 1\right\} \tag{6}$$

The amount of spectral sharpening is set by the parameter $\rho \geq 0$. As the quotient $T_k[n]/\tilde{e}_k[n]$ is restricted to values below 1, any $\rho > 0$ literally suppresses lower quotients.

The effect of spectral sharpening is demonstrated by knocking with knuckles on a wooden plate. Listen to the signal without and with spectral ACE (Snd. 1.1 and 1.2, respectively). Corresponding spectrograms are shown in Fig. 5a-b. Parameters have been set to values which work well for most signals: $\rho = 30$, $\sigma = 3$ ERB, and smoothing time constant $\tau = 7$ ms. It is apparent that the described algorithm effectively suppresses spectral troughs while leaving local maxima as narrowband regions with their original amplitude. In addition, the broadband background noise is reduced to some high-frequency artifacts of the recording which are now

clearly audible. A ρ larger than 30 does not seem to bring any benefit for spectral sharpening; the signal is already reduced to its local maxima. Additional contrast can be achieved by spectral dynamics expansion, as explained in the next section.

2.1.2. Spectral Dynamics Expansion

The goal of spectral dynamics expansion is to attenuate frequency bands with low energy while pulling those with high energy, above a certain threshold value, up to the running global maximum. In contrast to spectral sharpening, this approach should not attenuate broadband regions in the spectrum if they are prominent enough. On the downside, it will suppress even very prominent local maxima if they appear below the threshold.

Spectral dynamics processing is achieved by exponentiation of the magnitude spectrum — inspired by the simple algorithm originally proposed by Boers [14]. In our case, each envelope $u_k[n]$ is scaled with respect to the global maximum of all (smoothed) envelopes (see Eq. 7). As gain-factor, we use the quotient of the smoothed envelope $\tilde{u}_k[n]$ and a fraction of the instantaneous maximum of all smoothed envelopes ($\mu\tilde{u}_{max}$). The exponent $\beta \geq 0$ sets the amount of expansion; $0 < \mu \leq 1$ is the relative threshold. Gain is clipped at $\tilde{u}_{max}/\tilde{u}_k[n]$ so that $u_k[n]$ does not exceed the maximum of all sub-band envelopes.

$$v_k[n] = u_k[n] \cdot \min\left\{ \left(\frac{\tilde{u}_k[n]}{\mu\tilde{u}_{max}[n]} \right)^{\beta}, \; \frac{\tilde{u}_{max}[n]}{\tilde{u}_k[n]} \right\} \quad (7)$$

with the (instantaneous) global maximum

$$\tilde{u}_{max}[n] = \max_k \big\{ \tilde{u}_k[n] \big\}. \quad (8)$$

Listen again to the enhanced signal from the previous section (Snd. 1.2 / Fig. 5b). Additional contrast is achieved by feeding this signal into spectral dynamics expansion (Snd. 1.3 / Fig. 5c). Furthermore, the background noise is gone. The parameters have been set to $\mu = 0.8$ and an extreme value of $\beta = 8$, leading to a spectral gate where values below $\mu\tilde{u}_{max}[n]$ are almost completely suppressed while values above approach the global maximum.

Contrary to spectral sharpening, spectral dynamics expansion can also be used to exaggerate broadband regions in the spectrum. This is demonstrated in Snd. 2.1 and 2.2 with the recording of a vintage printing machine, with noise from a pneumatic system.

2.1.3. Decay Prolongation

Spectral resolution and pitch impression takes time. What if we gave listeners more time to perceive a sound by prolonging it through artificial decay? Such an effect could be achieved in a natural way via reverberation. Dombois and Eckel argue that reverberation might even be used to enhance audifications, as it facilitates discrimination between short transient sounds [26, p. 315]. Koumura and Furukawa examined the effect of reverberation on the identification of material via short impact sounds [27]. They found out that reverberation actually deteriorates material identification; however, after a short while, participants adapted to the reverberation and achieved similar identification rates as with the dry stimuli. It must be noted that the results varied greatly among participants. Furthermore, adaptation to reverberation during speech does not help to identify a following impact sound [28]. Such natural reverberation, of course, is not correlated to the stimulus itself, but just convolves it with an arbitrary impulse response. A completely "transparent" reverberation whose impulse response has a white magnitude spectrum might already lead to better results.

Yet another problem is the broadband spectrum of the transient sounds — any artificial reverberation will therefore mask succeeding parts completely with broadband noise. Even if the resonances are sharpened through spectral contrast enhancement as derived in Sec. 2.1, a short transient signal in a single sub-band still results in a broadband signal at the output. However, if artificial decay is applied to the individual sub-band envelopes, their bandwidths are reduced and more time is given to the listener to gain a pitch impression. The enhanced sub-band envelopes after lateral inhibition and exponentiation may still contain short spikes which are not visible in the spectrogram of Fig. 5b-c, but which would have a huge impact if the sub-band envelopes were decayed as they are. Therefore, the envelopes must be smoothed before decay prolongation. As this further smears the envelopes in time, we instead split them into a transient part and a decay part. Only the decay part receives decay prolongation; both are re-combined afterwards.

We first introduce two simple non-linear low-pass filters based on a leaky integrator. $\mathrm{env_a}$ has a smooth attack but instant decay, while $\mathrm{env_d}$ has a smooth decay but instant attack. $\mathrm{env_a}$ is given in Eq. 9 for an arbitrary input signal $x[n]$ and output signal $y[n]$. $\mathrm{env_d}$ follows the same equation, but with flipped direction of the inequality sign, leading to a naturally-sounding exponential decay.

$$y[n] = \begin{cases} (1-\alpha)|x[n]| + \alpha y[n-1], & |x[n]| < y[n-1] \\ |x_k[n]|, & \text{otherwise} \end{cases} \quad (9)$$

The amount of smoothing is set via the smoothing factor α. A more convenient parametrization can be achieved via time constant τ or -60 dB reverberation time T_{60}:

$$\alpha = \exp\left(-\frac{1}{\tau f_s}\right) = \exp\left(-\frac{\ln(1000)}{T_{60} f_s}\right), \quad (10)$$

where f_s is the sampling frequency.

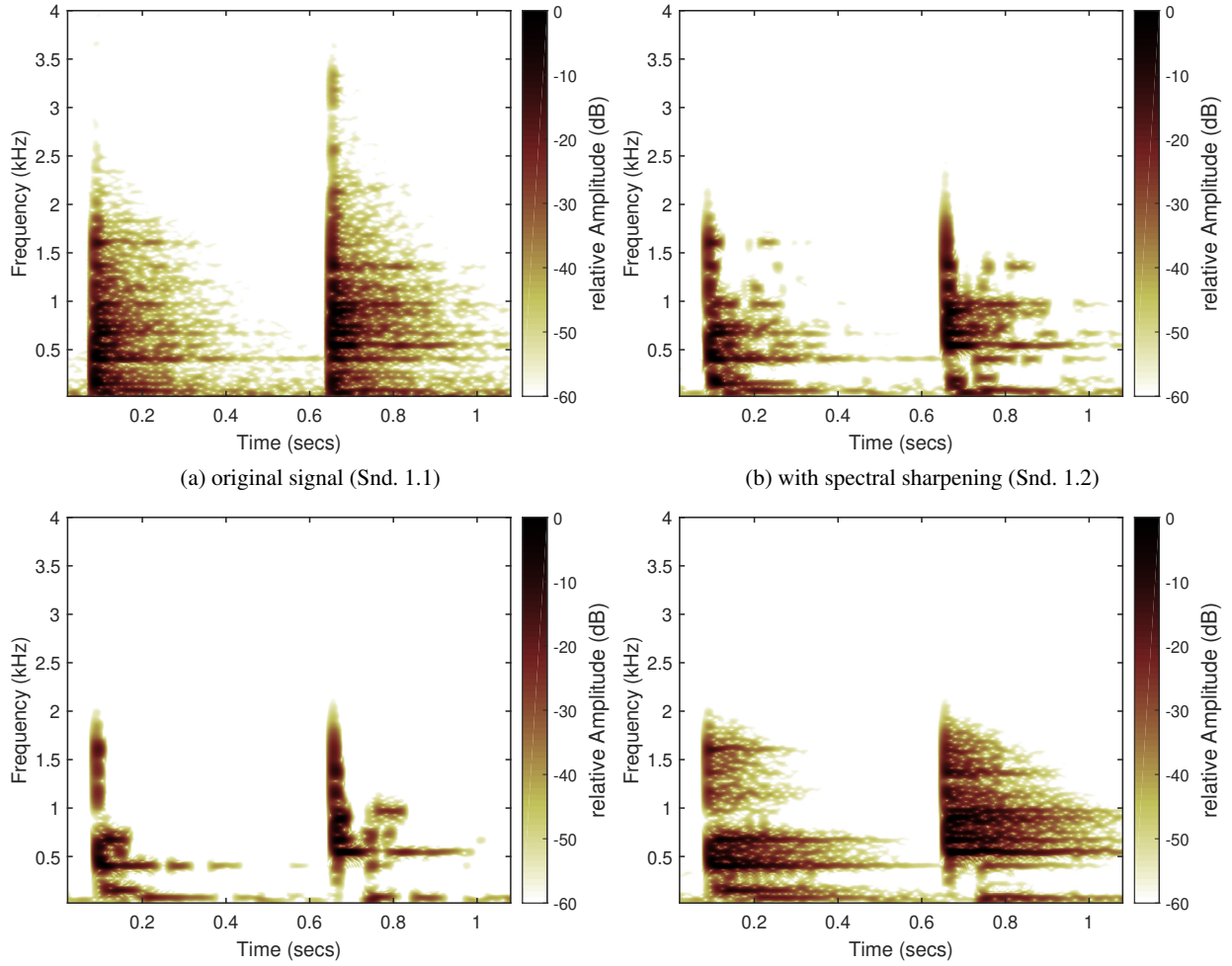
The envelope with smoothed attack $\mathrm{env_a}\{v_k[n]\}$ is fed to decay prolongation, while the residuum $(v_k[n] - \mathrm{env_a}\{v_k[n]\})$ containing only the attack part is added back to the result, leading to the output signal of decay prolongation $p_k[n]$:

$$p_k[n] = \mathrm{env_d}\Big\{\mathrm{env_a}\big\{v_k[n]\big\}\Big\} + v_k[n] - \mathrm{env_a}\big\{v_k[n]\big\}. \quad (11)$$

Due to the normalization with the original envelopes (Eq. 2) the decay is fed by intrinsic signal components of the sub-band signals in the relevant frequency region. In order to supply sufficient signal energy in the case of large SNR combined with long decay prolongation, a pink noise signal $\eta[n]$ is added to the input signal just before feeding it to the Gammatone filterbank (see block diagram in Fig. 4a); at a level below the threshold of hearing, but enough to synthesize literally infinite decay. As internal signal processing on any eligible platform offers at least 32 bit floating-point precision, a noise level of around -96 dBFS is more than enough.

A constant decay time over the whole frequency range leads to an unnatural amplification of high frequencies, as damping usually increases with frequency. We chose a rough approximation by setting T_{60} inversely proportional to the center frequency, but clipped below 1 kHz.

Sound example 1.3 and Fig. 5d show the effect of decay prolongation on the enhanced signal from Sec. 2.1.2 (Snd. 1.3 and Fig. 5 c). For this example, reverberation time T_{60} at 1 kHz was



(a) original signal (Snd. 1.1)

(b) with spectral sharpening (Snd. 1.2)

(c) with spectral dynamics expansion (Snd. 1.3)

(d) with decay prolongation (Snd. 1.4)

Figure 5: Spectrograms of a test sound in 4 conditions: (a) original recording, (b) with spectral sharpening, (c) with spectral sharpening and spectral dynamics expansion, (d) with spectral sharpening, spectral dynamics expansion, and decay prolongation.

set to 0.5 s. The time constant for transient separation was set to 7 ms. It is clearly visible and audible that relevant partials are significantly extended in time.

2.2. Temporal Contrast Enhancement

Temporal contrast enhancement is done for two reasons: (1) to make temporal structures in the sound more prominent, and (2) to compensate latency and time-smearing of the spectral contrast enhancement. Spectral ACE, as described above, always introduces some latency which is small at high frequencies but increases towards lower frequencies. This frequency-dependent group delay is acceptable for steady sounds, but it delays and smoothes any transient, transforming it to something similar to a down-chirp. Due to their broadband spectrum in combination with smoothed lateral inhibition, spectral ACE anyway effectively suppresses all transients. In order to preserve them, they must be detected as fast as possible from the input signal and mixed together with the output of spectral contrast enhancement and delay prolongation.

Transients are detected in real time by the same simple transient detection algorithm that has been used for decay prolongation (see Sec. 2.1.3). A 2nd-order high-pass filter with adjustable cutoff frequency makes the transient detection more sensitive to high-frequency content. $s_h[n]$ is the high-pass-filtered version of $s[n]$. The envelope $e_t[n]$ of the transient part of the signal is estimated via the difference of a slowly decaying envelope $e_{t,d}[n]$ and a slowly rising envelope $e_{t,a}[n]$:

$$e_t[n] = \max\left\{e_{t,d}[n] - e_{t,a}[n] - \nu\ ,\ \ 0\right\} \tag{12}$$

with threshold ν. Envelopes are computed via the two filters $\mathrm{env_d}$ and $\mathrm{env_a}$ that have been explained in Sec. 2.1.3 and Eqs. 9–10:

$$e_{t,d}[n] = \mathrm{env_d}\big\{s_h[n]\big\} \quad , \quad e_{t,a}[n] = \mathrm{env_a}\big\{e_{t,d}[n]\big\}\ . \tag{13}$$

The output signal of temporal ACE, $s_t[n]$, contains only the detected transients with their original amplitude:

$$s_t[n] = s[n] \cdot \frac{e_t[n]}{\mathrm{env_d}\big\{e_t[n]\big\}}\ . \tag{14}$$

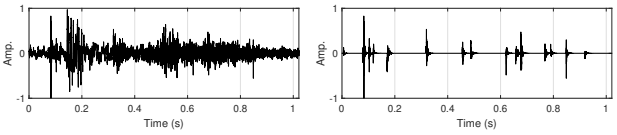


Figure 6: Waveform of the signal without (left, Snd. 2.1) and with temporal ACE (right, Snd. 2.3).

Setting the time constants $\tau_a = 3\,\text{ms}$ for env_a and $\tau_d = 7\,\text{ms}$ for env_a, seems to work well with most of the signals we tested. Threshold ν is adjusted dependent to the overall signal level.

In sound examples Snd. 1.2-1.4, the original transients are smoothed by spectral contrast enhancement. For that reason, the original transients are extracted (Snd. 1.5) and mixed to the enhanced signals. Sound examples 1.6-1.8 are the same as Snd. 1.2-1.4, respectively, but with restored transients.

In Fig. 6 and Snd. 2.3, the effect of temporal contrast enhancement is demonstrated with the machine recording from Snd. 2.1 It is clearly visible that, similar to spectral sharpening, local amplitude minima are attenuated while local amplitude maxima are retained. Note that the algorithm operates on the highpass-filtered version (cutoff frequency set to 4 kHz). The mechanic rattling thus becomes the prominent sound characteristic. A mix with the enhanced signal from spectral dynamics expansion (Snd. 2.2) leads to a spectrally and temporally enhanced signal (Snd. 2.4).

For temporal contrast enhancement, it makes no sense to apply dynamics expansion based on an absolute threshold as for spectral contrast enhancement via exponentiation — this would be a waveshaper, introducing unwanted distortion. The linear crossfade with the dry input signal actually serves as a control for the amplitude of the residuum signal between transients.

3. DISCUSSION

One might notice that the proposed ACE method does not explicitly include spectro-temporal contrast enhancement, e.g., temporal contrast enhancement on a sub-band level. Our hearing system does exactly that via contrast gain control in the auditory cortex, at timescales of about 100 ms [29]. Rabinowitz et al. define spectro-temporal contrast as "the variation in sound pressure in each frequency band, relative to the mean"; a model can be based on the standard deviation of recent sound pressure level [29]. One audible effect is that a harmonic partial which is omitted and then reintroduced may stand out perceptually for a short period of time [30]. While this is certainly a helpful feature, it must be noted that the main objective of such adaptive gain control is to compensate the very limited dynamic range of neurons. We found that spectro-temporal contrast is anyway strong with spectral contrast enhancement alone, e.g., through a possible edge effect in case of a missing partial. Even more so, if smoothing for lateral inhibition is bypassed, together with a large ρ, a clicking transient appears whenever there is a shift of spectral energy from one band to another. Due to the group delay of the filters, however, such a transient would exhibit latency that is unacceptable for short interaction sounds.

For continuous sounds where more latency can be tolerated, it might be interesting to exaggerate amplitude modulations on a sub-band level. For that goal we tried an algorithm which expands the sub-band envelopes individually while preserving their overall envelope trend [31]. While originally designed to exaggerate dissonances, it is capable to enhance also low-frequency amplitude modulations. At a closer look, however, similar results could be achieved by spectral ACE alone.

Concerning spectral contrast, both methods — spectral sharpening and spectral dynamics expansion — are essential. As soon as spectral sharpening has reached its limits (i.e., what is left are local maxima only), spectral dynamics expansion can add additional contrast by suppressing all local maxima below a certain threshold.

In a parallel configuration, spectral sharpening and spectral dynamics expansion can complement each other, producing a cartoonification of the sound. This may be illustrated by the example of human speech: By lateral inhibition, speech is basically reduced to fundamental frequency and formants; consonants are attenuated. While stops/plosives could be recovered via temporal contrast enhancement, sibilants are suppressed. Exponentiation maintains or even exaggerates consonants, including sibilants; however, it has a tendency to suppress formants, so that discrimination between vowels is lost. The solution might be a combination by taking the maximum of both outputs.

Temporal contrast enhancement as implemented here works similar to a transient shaper/designer for music production. The main difference is that we try not to exaggerate transients but to attenuate everything else. A dynamics expansion would conflict with the limited dynamic range of our hearing system, and would also produce an implausible amplification of the targeted interaction sounds. The mix of spectral and temporal ACE works well for these impact sounds sounds, but may produces quite disturbing results for more continuous stimuli such as speech.

4. CONCLUSIONS AND OUTLOOK

We introduced a new method for real-time auditory contrast enhancement, targeting at interactive applications where auditory feedback is used as part of a knowledge-making process. The method is split in two parts — spectral and temporal contrast enhancement — which can be used in parallel to focus on different auditory features. Spectral ACE is achieved in two ways which both are needed for different tasks. While the first approach is based on lateral inhibition and enhances spectral sharpness, the second enhances spectral dynamics via exponentiation. In the visual domain, these would refer to edge detection and contrast, respectively. Crucial for perceptibility of the enhanced sound is decay prolongation which provides a listener with additional time for pitch impression. Transient detection was found to be sufficient for temporal contrast enhancement. First results indicate that auditory contrast can be significantly enhanced by the proposed method.

The next step is to evaluate the multitude of parameters in order to find meaningful ranges and scalings, and ultimately reduce them to only a few intuitive controls. A parameter study is planned to find a compromise, achieving high auditory contrast while maintaining a certain degree of naturalness and plausibility of any auditory feedback. Participants will be rating the plausibility of observed interactions (audition vs. vision) through short video sequences, with different settings of ACE applied to the audio track. Recordings are taken from the Greatest Hits dataset [32], a collection of audio/video recordings of different kinds of objects and materials being hit with a drumstick.

It is further planned to evaluate the presented method concerning its primary target application: percussion. Contrary to the parameter study, interaction will be performed by the participants

themselves. The technical setup can be regarded as a special case of auditory augmentation, similar to the augmented table described in [1, 4]; however, with electronics not hidden but clearly visible, e.g., as a mic-through system. Participants will be asked to identify position and type of concealed physical manipulations (e.g., cavity or thickening) below the visible surface, via percussion with fingers or a hammer tool. Performance with ACE will be compared to the control condition without ACE; qualitative interviews should reveal further implications.

5. REFERENCES

[1] M. Weger, T. Hermann, and R. Höldrich, "Plausible auditory augmentation of physical interaction," in *ICAD*, 2018.

[2] T. Bovermann, R. Tünnermann, and T. Hermann, "Auditory Augmentation," *International Journal on Ambient Computing and Intelligence (IJACI)*, vol. 2, no. 2, pp. 27–41, 2010.

[3] K. Groß-Vogt, M. Weger, R. Höldrich, T. Hermann, T. Bovermann, and S. Reichmann, "Augmentation of an institute's kitchen: An ambient auditory display of electric power consumption," in *ICAD*, 2018.

[4] K. Groß-Vogt, M. Weger, and R. Höldrich, "Exploration of auditory augmentation in an interdisciplinary prototyping workshop," in *Forum Media Technology*, 2018.

[5] S. Papetti and F. Fontana, "Effects of audio-tactile floor augmentation on perception and action during walking: Preliminary results," in *SMC*, 2012, pp. 17–22.

[6] E. Furfaro, F. Bevilacqua, N. Berthouze, and A. Tajadura-Jimenez, "Sonification of virtual and real surface tapping: evaluation of behavior changes, surface perception and emotional indices," *IEEE MultiMedia*, 2015.

[7] J. Maculewicz, C. Erkut, and S. Serafin, "An investigation on the influence of soundscapes and footstep sounds in affecting preferred walking pace," in *ICAD*, 2015.

[8] A. Supper and K. Bijsterveld, "Sounds convincing: Modes of listening and sonic skills in knowledge making," *Interdisciplinary Science Reviews*, vol. 40, no. 2, pp. 124–144, 2015.

[9] P. Y. Ertel, M. Lawrence, and W. Song, "Stethoscope acoustics and the engineer: Concepts and problems," *Journal of the AES*, vol. 19, no. 3, pp. 182–186, 1971.

[10] T. Hermann and M. Weger, "Data-driven auditory contrast enhancement for everyday sounds and sonifications," in *ICAD*, Newcastle, U.K., 2019.

[11] P. Susini, N. Misdariis, G. Lemaitre, and O. Houix, "Naturalness influences the perceived usability and pleasantness of an interface's sonic feedback," *Journal on Multimodal User Interfaces*, vol. 5, no. 3-4, pp. 175–186, 2012.

[12] A. P. McPherson, R. H. Jack, G. Moro, *et al.*, "Action-sound latency: Are our tools fast enough?" in *NIME*, 2016.

[13] J. Yang, F.-L. Luo, and A. Nehorai, "Spectral contrast enhancement: Algorithms and comparisons," *Speech Communication*, vol. 39, no. 1-2, pp. 33–46, 2003.

[14] P. Boers, "Formant enhancement of speech for listeners with sensorineural hearing loss," *IPO annual progress report*, vol. 15, pp. 21–28, 1980.

[15] M. A. Stone and B. C. Moore, "Spectral feature enhancement for people with sensorineural hearing impairment: Effects on speech intelligibility and quality," *Journal of rehabilitation research and development*, vol. 29, no. 2, pp. 39–56, 1992.

[16] T. Baer, B. C. Moore, and S. Gatehouse, "Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: Effects on intelligibility, quality, and response times," *Journal of rehabilitation research and development*, vol. 30, pp. 49–49, 1993.

[17] B. C. Moore and B. R. Glasberg, "A revision of zwicker's loudness model," *Acta Acustica united with Acustica*, vol. 82, no. 2, pp. 335–345, 1996.

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

[19] C. Pantev, H. Okamoto, B. Ross, W. Stoll, E. Ciurlia-Guy, R. Kakigi, and T. Kubo, "Lateral inhibition and habituation of the human auditory cortex," *European Journal of Neuroscience*, vol. 19, no. 8, pp. 2337–2344, 2004.

[20] T. Houtgast, "Psychophysical evidence for lateral inhibition in hearing," *JASA*, vol. 51, no. 6B, pp. 1885–1894, 1972.

[21] S. Coren, C. Porac, D. J. Aks, and K. Morikawa, "A method to assess the relative contribution of lateral inhibition to the magnitude of visual-geometric illusions," *Perception & psychophysics*, vol. 43, no. 6, pp. 551–558, 1988.

[22] G. Békésy, "Lateral inhibition of heat sensations on the skin," *Applied physiology*, vol. 17, no. 6, pp. 1003–1008, 1962.

[23] R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," in *Meeting of the IOC Speech Group on Auditory Modelling at RSRE*, vol. 2, no. 7, 1987.

[24] V. Hohmann, "Frequency analysis and synthesis using a gammatone filterbank," *Acta Acustica united with Acustica*, vol. 88, no. 3, pp. 433–442, 2002.

[25] M. Noisternig, "Wechselwirkung von lautsprecher-mikrofon anordnungen in fahrzeugen," Dissertation, Graz University of Music and Performing Arts, 2017.

[26] T. Hermann, A. Hunt, and J. G. Neuhoff, Eds., *The sonification handbook*. Logos Verlag Berlin, Germany, 2011.

[27] T. Koumura and S. Furukawa, "Context-dependent effect of reverberation on material perception from impact sound," *Scientific reports*, vol. 7, no. 1, p. 16455, 2017.

[28] ——, "Do speech contexts induce constancy of material perception based on impact sound under reverberation?" *Acta Acustica u. w. Acustica*, vol. 104, no. 5, pp. 796–799, 2018.

[29] N. C. Rabinowitz, B. D. Willmore, J. W. Schnupp, and A. J. King, "Contrast gain control in auditory cortex," *Neuron*, vol. 70, no. 6, pp. 1178–1191, 2011.

[30] Q. Summerfield, A. Sidwell, and T. Nelson, "Auditory enhancement of changes in spectral amplitude," *JASA*, vol. 81, no. 3, pp. 700–708, 1987.

[31] M. Hoffman and P. Cook, "Real-time dissonancizers: Two dissonance-augmenting audio effects," in *DAFx*, 2008.

[32] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually indicated sounds," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2405–2413.